CHAPTER 3

Integration

SECTION 15. THE INTEGRAL

Expected values of simple random variables and Riemann integrals of continuous functions can be brought together with other related concepts under a general theory of integration, and this theory is the subject of the present chapter.

Definition

Throughout this section, f, g, and so on will denote real measurable functions, the values $\pm\infty$ allowed, on a measure space $(\Omega, \mathscr{F}, \mu)$.[†] The object is to define and study the definite integral

$$\int f d\mu = \int_\Omega f(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\mu(d\omega).$$

Suppose first that f is nonnegative. For each finite decomposition $\{A_i\}$ of Ω into $\mathscr{F}$-sets, consider the sum

$$\sum_i \left[\inf_{\omega \in A_i} f(\omega)\right] \mu(A_i). \tag{15.1}$$

In computing the products here, the conventions about infinity are

$$\begin{aligned} &0 \cdot \infty = \infty \cdot 0 = 0, \\ &x \cdot \infty = \infty \cdot x = \infty \qquad \text{if } 0 < x < \infty, \\ &\infty \cdot \infty = \infty. \end{aligned} \tag{15.2}$$

[†]Although the definitions (15.3) and (15.6) apply even if f is not measurable $\mathscr{F}$, the proofs of most theorems about integration do use the assumption of measurability in one way or another. For the role of measurability, and for alternative definitions of the integral, see the problems.

The reasons for these conventions will become clear later. Also in force are the conventions of Section 10 for sums and limits involving infinity; see (10.3) and (10.4). If A_i is empty, the infimum in (15.1) is by the standard convention ∞; but then $\mu(A_i) = 0$, so that by the convention (15.2), this term makes no contribution to the sum (15.1).

The integral of f is defined as the supremum of the sums (15.1):

$$\int f d\mu = \sup \sum_i \left[\inf_{\omega \in A_i} f(\omega) \right] \mu(A_i). \tag{15.3}$$

The supremum here extends over all finite decompositions $\{A_i\}$ of Ω into $\mathcal{F}$-sets.

For general f, consider its *positive part*,

$$f^+(\omega) = \begin{cases} f(\omega) & \text{if } 0 \le f(\omega) \le \infty, \\ 0 & \text{if } -\infty \le f(\omega) \le 0 \end{cases} \tag{15.4}$$

and its *negative part*,

$$f^-(\omega) = \begin{cases} -f(\omega) & \text{if } -\infty \le f(\omega) \le 0, \\ 0 & \text{if } 0 \le f(\omega) \le \infty. \end{cases} \tag{15.5}$$

These functions are nonnegative and measurable, and $f = f^+ - f^-$. The general integral is defined by

$$\int f d\mu = \int f^+ \, d\mu - \int f^- \, d\mu, \tag{15.6}$$

unless $\int f^+ \, d\mu = \int f^- \, d\mu = \infty$, in which case f has no integral.

If $\int f^+ \, d\mu$ and $\int f^- \, d\mu$ are both finite, then f is *integrable*, or integrable μ, or summable, and has (15.6) as its *definite integral*. If $\int f^+ \, d\mu = \infty$ and $\int f^- \, d\mu < \infty$, then f is not integrable but in accordance with (15.6) is assigned ∞ as its definite integral. Similarly, if $\int f^+ \, d\mu < \infty$ and $\int f^- \, d\mu = \infty$, then f is not integrable but has definite integral $-\infty$. Note that f can have a definite integral without being integrable; it fails to have a definite integral if and only if its positive and negative parts both have infinite integrals.

The really important case of (15.6) is that in which $\int f^+ \, d\mu$ and $\int f^- \, d\mu$ are both finite. Allowing infinite integrals is a convention that simplifies the statements of various theorems, especially theorems involving nonnegative functions. Note that (15.6) is defined unless it involves "$\infty - \infty$"; if one term on the right is ∞ and the other is a finite real x, the difference is defined by the conventions $\infty - x = \infty$ and $x - \infty = -\infty$.

The extension of the integral from the nonnegative case to the general case is consistent: (15.6) agrees with (15.3) if f is nonnegative, because then $f^- \equiv 0$.

Nonnegative Functions

It is convenient first to analyze nonnegative functions.

Theorem 15.1. (i) *If $f = \sum_i x_i I_{A_i}$ is a nonnegative simple function, $\{A_i\}$ being a finite decomposition of Ω into $\mathcal{F}$-sets, then $\int f\,d\mu = \sum_i x_i \mu(A_i)$.*

(ii) *If $0 \le f(\omega) \le g(\omega)$ for all ω, then $\int f\,d\mu \le \int g\,d\mu$.*

(iii) *If $0 \le f_n(\omega) \uparrow f(\omega)$ for all ω, then $0 \le \int f_n\,d\mu \uparrow \int f\,d\mu$.*

(iv) *For nonnegative functions f and g and nonnegative constants α and β, $\int(\alpha f + \beta g)\,d\mu = \alpha \int f\,d\mu + \beta \int g\,d\mu$.*

In part (iii) the essential point is that $\int f\,d\mu = \lim_n \int f_n\,d\mu$, and it is important to understand that both sides of this equation may be ∞. If $f_n = I_{A_n}$ and $f = I_A$, where $A_n \uparrow A$, the conclusion is that μ is continuous from below (Theorem 10.2(i)): $\lim_n \mu(A_n) = \mu(A)$; this equation often takes the form $\infty = \infty$.

PROOF OF (i). Let $\{B_j\}$ be a finite decomposition of Ω and let β_j be the infimum of f over B_j. If $A_i \cap B_j \ne \varnothing$, then $\beta_j \le x_i$; therefore, $\sum_j \beta_j \mu(B_j) = \sum_{ij} \beta_j \mu((A_i \cap B_j)) \le \sum_{ij} x_i \mu(A_i \cap B_j) = \sum_i x_i \mu(A_i)$. On the other hand, there is equality here if $\{B_j\}$ coincides with $\{A_i\}$. ■

PROOF OF (ii). The sums (15.1) obviously do not decrease if f is replaced by g. ■

PROOF OF (iii). By (ii) the sequence $\int f_n\,d\mu$ is nondecreasing and bounded above by $\int f\,d\mu$. It therefore suffices to show that $\int f\,d\mu \le \lim_n \int f_n\,d\mu$, or that

$$\lim_n \int f_n\,d\mu \ge S = \sum_{i=1}^m v_i \mu(A_i) \tag{15.7}$$

if $A_1, \ldots, A_m$ is any decomposition of Ω into $\mathcal{F}$-sets and $v_i = \inf_{\omega \in A_i} f(\omega)$.

In order to see the essential idea of the proof, which is quite simple, suppose first that S is finite and all the v_i and $\mu(A_i)$ are positive and finite. Fix an ϵ that is positive and less than each v_i, and put $A_{in} = [\omega \in A_i: f_n(\omega) > v_i - \epsilon]$. Since $f_n \uparrow f$, $A_{in} \uparrow A_i$. Decompose Ω into $A_{1n}, \ldots, A_{mn}$ and the complement of their union, and observe that, since μ is continuous from below,

$$\begin{aligned} \int f_n\,d\mu &\ge \sum_{i=1}^m (v_i - \epsilon)\mu(A_{in}) \to \sum_{i=1}^m (v_i - \epsilon)\mu(A_i) \\ &= S - \epsilon \sum_{i=1}^m \mu(A_i). \end{aligned} \tag{15.8}$$

Since the $\mu(A_i)$ are all finite, letting $\epsilon \to 0$ gives (15.7).

Now suppose only that S is finite. Each product $v_i\mu(A_i)$ is then finite; suppose it is positive for $i \le m_0$ and 0 for $i > m_0$. (Here $m_0 \le m$; if the product is 0 for all i, then $S = 0$ and (15.7) is trivial.) Now v_i and $\mu(A_i)$ are positive and finite for $i \le m_0$ (one or the other may be ∞ for $i > m_0$). Define A_{in} as before, but only for $i \le m_0$. This time decompose Ω into $A_{1n}, \ldots, A_{m_0 n}$ and the complement of their union. Replace m by m_0 in (15.8) and complete the proof as before.

Finally, suppose that $S = \infty$. Then $v_{i_0}\mu(A_{i_0}) = \infty$ for some i_0, so that v_{i_0} and $\mu(A_{i_0})$ are both positive and at least one is ∞. Suppose $0 < x < v_{i_0} \le \infty$ and $0 < y < \mu(A_{i_0}) \le \infty$, and put $A_{i_0 n} = [\omega \in A_{i_0}\colon f_n(\omega) > x]$. From $f_n \uparrow f$ follows $A_{i_0 n} \uparrow A_{i_0}$; hence $\mu(A_{i_0 n}) > y$ for n exceeding some n_0. But then (decompose Ω into $A_{i_0 n}$ and its complement) $\int f_n\,d\mu \ge x\mu(A_{i_0 n}) \ge xy$ for $n > n_0$, and therefore $\lim_n \int f_n\,d\mu \ge xy$. If $v_{i_0} = \infty$, let $x \to \infty$, and if $\mu(A_{i_0}) = \infty$, let $y \to \infty$. In either case (15.7) follows: $\lim_n \int f_n\,d\mu = \infty$. ■

Proof of (iv). Suppose at first that $f = \sum_i x_i I_{A_i}$ and $g = \sum_j y_j I_{B_j}$ are simple. Then $\alpha f + \beta g = \sum_{ij}(\alpha x_i + \beta y_j) I_{A_i \cap B_j}$, and so

$$\int (\alpha f + \beta g)\,d\mu = \sum_{ij} (\alpha x_i + \beta y_j)\mu(A_i \cap B_j)$$

$$= \alpha \sum_i x_i \mu(A_i) + \beta \sum_j y_j \mu(B_j) = \alpha \int f\,d\mu + \beta \int g\,d\mu.$$

Note that the argument is valid if some of α, β, x_i, y_j are infinite. Apart from this possibility, the ideas are as in the proof of (5.21).

For general nonnegative f and g, there exist by Theorem 13.5 simple functions f_n and g_n such that $0 \le f_n \uparrow f$ and $0 \le g_n \uparrow g$. But then $0 \le \alpha f_n + \beta g_n \uparrow \alpha f + \beta g$ and $\int(\alpha f_n + \beta g_n)\,d\mu = \alpha \int f_n\,d\mu + \beta \int g_n\,d\mu$, so that (iv) follows from (iii). ■

By part (i) of Theorem 15.1, the expected values of simple random variables in Chapter 1 are integrals: $E[X] = \int X(\omega) P(d\omega)$. This also covers the step functions in Section 1 (see (1.6)). The relation between the Riemann integral and the integral as defined here will be studied in Section 17.

Example 15.1. Consider the line $(R^1, \mathscr{R}^1, \lambda)$ with Lebesgue measure. Suppose that $-\infty < a_0 \le a_1 \le \cdots \le a_m < \infty$, and let f be the function with nonnegative value x_i on $(a_{i-1}, a_i]$, $i = 1, \ldots, m$, and value 0 on $(-\infty, a_0]$ and (a_m, ∞). By part (i) of Theorem 15.1, $\int f\,d\lambda = \sum_{i=1}^m x_i(a_i - a_{i-1})$ because of the convention $0 \cdot \infty = 0$—see (15.2). If the "area under the curve" to the left of a_0 and to the right of a_m is to be 0, this convention is inevitable. From $\infty \cdot 0 = 0$ it follows that $\int f\,d\lambda = 0$ if f is ∞ at a single point (say) and 0 elsewhere.

If $f = I_{(a,\infty)}$, the area-under-the-curve point of view makes $\int f\,d\mu = \infty$ natural. Hence the second convention in (15.2), which also requires that the integral be infinite if f is ∞ on a nonempty interval and 0 elsewhere. ■

Recall that *almost everywhere* means outside a set of measure 0.

Theorem 15.2. *Suppose that f and g are nonnegative.*

(i) *If $f = 0$ almost everywhere, then $\int f\,d\mu = 0$.*
(ii) *If $\mu[\omega\colon f(\omega) > 0] > 0$, then $\int f\,d\mu > 0$.*
(iii) *If $\int f\,d\mu < \infty$, then $f < \infty$ almost everywhere.*
(iv) *If $f \le g$ almost everywhere, then $\int f\,d\mu \le \int g\,d\mu$.*
(v) *If $f = g$ almost everywhere, then $\int f\,d\mu = \int g\,d\mu$.*

PROOF. Suppose that $f = 0$ almost everywhere. If A_i meets $[\omega\colon f(\omega) = 0]$, then the infimum in (15.1) is 0; otherwise, $\mu(A_i) = 0$. Hence each sum (15.1) is 0, and (i) follows.

If $A_\epsilon = [\omega\colon f(\omega) \ge \epsilon]$, then $A_\epsilon \uparrow [\omega\colon f(\omega) > 0]$ as $\epsilon \downarrow 0$, so that under the hypothesis of (ii) there is a positive ϵ for which $\mu(A_\epsilon) > 0$. Decomposing Ω into A_ϵ and its complement shows that $\int f\,d\mu \ge \epsilon\mu(A_\epsilon) > 0$.

If $\mu[f = \infty] > 0$, decompose Ω into $[f = \infty]$ and its complement: $\int f\,d\mu \ge \infty \cdot \mu[f = \infty] = \infty$ by the conventions. Hence (iii).

To prove (iv), let $G = [f \le g]$. For any finite decomposition $\{A_1, \ldots, A_m\}$ of Ω,

$$\sum \left[\inf_{A_i} f\right] \mu(A_i) = \sum \left[\inf_{A_i} f\right] \mu(A_i \cap G) \le \sum \left[\inf_{A_i \cap G} f\right] \mu(A_i \cap G)$$

$$\le \sum \left[\inf_{A_i \cap G} g\right] \mu(A_i \cap G) \le \int g\,d\mu,$$

where the last inequality comes from a consideration of the decomposition $A_1 \cap G, \ldots, A_m \cap G, G^c$. This proves (iv), and (v) follows immediately. ■

Suppose that $f = g$ almost everywhere, where f and g need not be nonnegative. If f has a definite integral, then since $f^+ = g^+$ and $f^- = g^-$ almost everywhere, it follows by Theorem 15.2(v) that g also has a definite integral and $\int f\,d\mu = \int g\,d\mu$.

Uniqueness

Although there are various ways to frame the definition of the integral, they are all equivalent—they all assign the same value to $\int f\,d\mu$. This is because the integral is uniquely determined by certain simple properties it is natural to require of it.

It is natural to want the integral to have properties (i) and (iii) of Theorem 15.1. But these uniquely determine the integral for nonnegative functions: For f nonnegative, there exist by Theorem 13.5 simple functions f_n such that $0 \le f_n \uparrow f$; by (iii), $\int f\,d\mu$ must be $\lim_n \int f_n\,d\mu$, and (i) determines the value of each $\int f_n\,d\mu$.

Property (i) can itself be derived from (iv) (linearity) together with the assumption that $\int I_A\,d\mu = \mu(A)$ for indicators I_A: $\int(\sum_i x_i I_{A_i})\,d\mu = \sum_i x_i \int I_{A_i}\,d\mu = \sum_i x_i \mu(A_i)$.

If (iv) of Theorem 15.1 is to persist when the integral is extended beyond the class of nonnegative functions, $\int f\,d\mu$ must be $\int(f^+ - f^-)\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu$, which makes the definition (15.6) inevitable.

PROBLEMS

These problems outline alternative definitions of the integral and clarify the role measurability plays. Call (15.3) the *lower integral*, and write it as

$$\int_* f\,d\mu = \sup \sum_i \left[\inf_{\omega\in A_i} f(\omega)\right]\mu(A_i) \tag{15.9}$$

to distinguish it from the *upper integral*

$$\int^* f\,d\mu = \inf \sum_i \left[\sup_{\omega\in A_i} f(\omega)\right]\mu(A_i). \tag{15.10}$$

The infimum in (15.10), like the supremum in (15.9), extends over all finite partitions $\{A_i\}$ of Ω into $\mathscr{F}$-sets.

15.1. Suppose that f is measurable and nonnegative. Show that $\int^* f\,d\mu = \infty$ if $\mu[\omega\colon f(\omega) > 0] = \infty$ or if $\mu[\omega\colon f(\omega) > a] > 0$ for all a.

There are many functions familiar from calculus that ought to be integrable but are of the types in the preceding problem and hence have infinite upper integral. Examples are $x^{-2}I_{(1,\infty)}(x)$ and $x^{-1/2}I_{(0,1)}(x)$. Therefore, (15.10) is inappropriate as a definition of $\int f\,d\mu$ for nonnegative f. The only problem with (15.10), however, is that it treats infinity the wrong way. To see this, and to focus on essentials, assume that $\mu(\Omega) < \infty$ and that f is bounded, although not necessarily nonnegative or measurable $\mathscr{F}$.

15.2. ↑ **(a)** Show that

$$\sum_i \left[\inf_{\omega\in A_i} f(\omega)\right]\mu(A_i) \le \sum_j \left[\inf_{\omega\in B_j} f(\omega)\right]\mu(B_j)$$

if $\{B_j\}$ refines $\{A_i\}$. Prove a dual relation for the sums in (15.10) and conclude that

$$\int_* f\,d\mu \le \int^* f\,d\mu. \tag{15.11}$$

(b) Now assume that f is measurable $\mathscr{F}$ and let M be a bound for $|f|$. Consider the partition $A_i = [\omega\colon i\epsilon < f(\omega) \le (i+1)\epsilon]$, where i ranges from $-N$

to N and N is large enough that $N\epsilon > M$. Show that

$$\sum_i \left[\sup_{\omega\in A_i} f(\omega)\right]\mu(A_i) - \sum_i \left[\inf_{\omega\in A_i} f(\omega)\right]\mu(A_i) \le \epsilon\mu(\Omega).$$

Conclude that

$$\int_* f\,d\mu = \int^* f\,d\mu. \tag{15.12}$$

To define the integral as the common value in (15.12) is the *Darboux–Young* approach. The advantage of (15.3) as a definition is that (in the nonnegative case) it applies at once to unbounded f and infinite μ.

15.3. 3.2 15.2↑ For $A \subset \Omega$, define $\mu^*(A)$ and $\mu_*(A)$ by (3.9) and (3.10) with μ in place of P. Show that $\int^* I_A\,d\mu = \mu^*(A)$ and $\int_* I_A\,d\mu = \mu_*(A)$ for every A. Therefore, (15.12) can fail if f is not measurable $\mathscr{F}$. (Where was measurability used in the proof of (15.12)?)

The definitions (15.3) and (15.6) always make formal sense (for finite $\mu(\Omega)$ and $\sup|f|$), but they are reasonable—accord with intuition—only if (15.12) holds. Under what conditions *does* it hold?

15.4. 10.5 15.3↑ (**a**) Suppose of f that *there exist an $\mathscr{F}$-set A and a function g, measurable $\mathscr{F}$, such that $\mu(A) = 0$ and $[f \ne g] \subset A$.* This is the same thing as assuming that $\mu^*[f \ne g] = 0$, or assuming that f is measurable with respect to $\mathscr{F}$ completed with respect to μ. Show that (15.12) holds.

(**b**) Show that if (15.12) holds, then so does the italicized condition in part (a).

Rather than assume that f is measurable $\mathscr{F}$, one can assume that it satisfies the italicized condition in Problem 15.4(a)—which in case $(\Omega, \mathscr{F}, \mu)$ is complete is the same thing anyway. For the next three problems, assume that $\mu(\Omega) < \infty$ and that f is measurable $\mathscr{F}$ and bounded.

15.5. ↑ Show that for positive ϵ there exists a finite partition $\{A_i\}$ such that, if $\{B_j\}$ is any finer partition and $\omega_j \in B_j$, then

$$\left|\int f\,d\mu - \sum_j f(\omega_j)\mu(B_j)\right| < \epsilon.$$

15.6. ↑ Show that

$$\int f\,d\mu = \lim_n \sum_{|k|\le n2^n} \frac{k-1}{2^n}\mu\left[\omega\colon \frac{k-1}{2^n} \le f(\omega) < \frac{k}{2^n}\right].$$

The limit on the right here is *Lebesgue's* definition of the integral.

15.7. ↑ Suppose that the integral is *defined* for simple nonnegative functions by $\int(\sum_i x_i I_{A_i})\,d\mu = \sum_i x_i \mu(A_i)$. Suppose that f_n and g_n are simple and nondecreasing and have a common limit: $0 \le f_n \uparrow f$ and $0 \le g_n \uparrow f$. Adapt the arguments used to prove Theorem 15.1(iii) and show that $\lim_n \int f_n\,d\mu = \lim_n \int g_n\,d\mu$. Thus, in the nonnegative case, $\int f\,d\mu$ can (Theorem 13.5) consistently be *defined* as $\lim_n \int f_n\,d\mu$ for simple functions for which $0 \le f_n \uparrow f$.

SECTION 16. PROPERTIES OF THE INTEGRAL

Equalities and Inequalities

By definition, the requirement for integrability of f is that $\int f^+\,d\mu$ and $\int f^-\,d\mu$ both be finite, which is the same as the requirement that $\int f^+\,d\mu + \int f^-\,d\mu < \infty$ and hence is the same as the requirement that $\int(f^+ + f^-)\,d\mu < \infty$ (Theorem 15.1(iv)). Since $f^+ + f^- = |f|$, f is integrable if and only if

$$\int |f|\,d\mu < \infty. \tag{16.1}$$

It follows that if $|f| \le |g|$ almost everywhere and g is integrable, then f is integrable as well. If $\mu(\Omega) < \infty$, a bounded f is integrable.

Theorem 16.1. (i) *Monotonicity*: *If f and g are integrable and $f \le g$ almost everywhere, then*

$$\int f\,d\mu \le \int g\,d\mu. \tag{16.2}$$

(ii) *Linearity*: *If f and g are integrable and α, β are finite real numbers, then $\alpha f + \beta g$ is integrable and*

$$\int(\alpha f + \beta g)\,d\mu = \alpha \int f\,d\mu + \beta \int g\,d\mu. \tag{16.3}$$

PROOF OF (i). For nonnegative f and g such that $f \le g$ almost everywhere, (16.2) follows by Theorem 15.2(iv). And for general integrable f and g, if $f \le g$ almost everywhere, then $f^+ \le g^+$ and $f^- \ge g^-$ almost everywhere, and so (16.2) follows by the definition (15.6). ■

PROOF OF (ii). First, $\alpha f + \beta g$ is integrable because, by Theorem 15.1,

$$\int |\alpha f + \beta g|\,d\mu \le \int(|\alpha| \cdot |f| + |\beta| \cdot |g|)\,d\mu$$

$$= |\alpha| \int |f|\,d\mu + |\beta| \int |g|\,d\mu < \infty.$$

By Theorem 15.1(iv) and the definition (15.6), $\int(\alpha f)\,d\mu = \alpha\int f\,d\mu$—consider separately the cases $\alpha \geq 0$ and $\alpha < 0$. Therefore, it is enough to check (16.3) for the case $\alpha = \beta = 1$. By definition, $(f+g)^+ - (f+g)^- = f+g = f^+ - f^- + g^+ - g^-$ and therefore $(f+g)^+ + f^- + g^- = (f+g)^- + f^+ + g^+$. All these functions being nonnegative, $\int(f+g)^+\,d\mu + \int f^-\,d\mu + \int g^-\,d\mu = \int(f+g)^-\,d\mu + \int f^+\,d\mu + \int g^+\,d\mu$, which can be rearranged to give $\int(f+g)^+\,d\mu - \int(f+g)^-\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu + \int g^+\,d\mu - \int g^-\,d\mu$. But this reduces to (16.3). ■

Since $-|f| \leq f \leq |f|$, it follows by Theorem 16.1 that

$$\left|\int f\,d\mu\right| \leq \int |f|\,d\mu \tag{16.4}$$

for integrable f. Applying this to integrable f and g gives

$$\left|\int f\,d\mu - \int g\,d\mu\right| \leq \int |f-g|\,d\mu. \tag{16.5}$$

Example 16.1. Suppose that Ω is countable, that $\mathcal{F}$ consists of all the subsets of Ω, and that μ is counting measure: each singleton has measure 1. To be definite, take $\Omega = \{1, 2, \ldots\}$. A function is then a sequence $x_1, x_2, \ldots$. If x_{nm} is x_m or 0 as $m \leq n$ or $m > n$, the function corresponding to $x_{n1}, x_{n2}, \ldots$ has integral $\sum_{m=1}^{n} x_m$ by Theorem 15.1(i) (consider the decomposition $\{1\}, \ldots, \{n\}, \{n+1, n+2, \ldots\}$). It follows by Theorem 15.1(iii) that in the nonnegative case the integral of the function given by $\{x_m\}$ is the sum $\sum_m x_m$ (finite or infinite) of the corresponding infinite series. In the general case the function is integrable if and only if $\sum_{m=1}^{\infty} |x_m|$ is a convergent infinite series, in which case the integral is $\sum_{m=1}^{\infty} x_m^+ - \sum_{m=1}^{\infty} x_m^-$.

The function $x_m = (-1)^{m+1} m^{-1}$ is not integrable by this definition and even fails to have a definite integral, since $\sum_{m=1}^{\infty} x_m^+ = \sum_{m=1}^{\infty} x_m^- = \infty$. This invites comparison with the ordinary theory of infinite series, according to which the alternating harmonic series does converge in the sense that $\lim_M \sum_{m=1}^{M} (-1)^{m+1} m^{-1} = \log 2$. But since this says that the sum of the *first* M terms has a limit, it requires that the elements of the space Ω be ordered. If Ω consists not of the positive integers but, say, of the integer lattice points in 3-space, it has no canonical linear ordering. And if $\sum_m x_m$ is to have the same finite value no matter what the order of summation, the series must be absolutely convergent.† This helps to explain why f is defined to be integrable only if $\int f^+\,d\mu$ and $\int f^-\,d\mu$ are both finite. ■

Example 16.2. In connection with Example 15.1, consider the function $f = 3I_{(a,\infty)} - 2I_{(-\infty,a)}$. There is no natural value for $\int f\,d\lambda$ (it is "$\infty - \infty$"), and none is assigned by the definition.

†RUDIN$_1$, p. 76.

If a function f is bounded on bounded intervals, then each function $f_n = fI_{(-n,n)}$ is integrable with respect to λ. Since $f = \lim_n f_n$, the limit of $\int f_n \, d\lambda$, if it exists, is sometimes called the "principal value" of the integral of f. Although it is natural for some purposes to integrate symmetrically about the origin, this is not the right definition of the integral in the context of general measure theory. The functions $g_n = fI_{(-n,n+1)}$ for example also converge to f, and $\int g_n \, d\lambda$ may have some other limit, or none at all; $f(x) = x$ is a case in point. There is no general reason why f_n should take precedence over g_n.

As in the preceding example, $f = \sum_{k=1}^{\infty} (-1)^k k^{-1} I_{(k,k+1)}$ has no integral, even though the $\int f_n \, d\lambda$ above converge. ■

Integration to the Limit

The first result, the *monotone convergence theorem*, essentially restates Theorem 15.1(iii).

Theorem 16.2. *If $0 \le f_n \uparrow f$ almost everywhere, then $\int f_n \, d\mu \uparrow \int f \, d\mu$.*

PROOF. If $0 \le f_n \uparrow f$ on a set A with $\mu(A^c) = 0$, then $0 \le f_n I_A \uparrow f I_A$ holds everywhere, and it follows by Theorem 15.1(iii) and the remark following Theorem 15.2 that $\int f_n \, d\mu = \int f_n I_A \, d\mu \uparrow \int f I_A \, d\mu = \int f \, d\mu$. ■

As the functions in Theorem 16.2 are nonnegative almost everywhere, all the integrals exist. The conclusion of the theorem is that $\lim_n \int f_n \, d\mu$ and $\int f \, d\mu$ are both infinite or both finite and in the latter case are equal.

Example 16.3. Consider the space $\{1, 2, \ldots\}$ together with counting measure, as in Example 16.1. If for each m one has $0 \le x_{nm} \uparrow x_m$ as $n \to \infty$, then $\lim_n \sum_m x_{nm} = \sum_m x_m$, a standard result about infinite series. ■

Example 16.4. If μ is a measure on $\mathcal{F}$, and $\mathcal{F}_0$ is a σ-field contained in $\mathcal{F}$, then the restriction μ_0 of μ to $\mathcal{F}_0$ is another measure (Example 10.4). If $f = I_A$ and $A \in \mathcal{F}_0$, then

$$\int f \, d\mu = \int f \, d\mu_0,$$

the common value being $\mu(A) = \mu_0(A)$. The same is true by linearity for nonnegative simple functions measurable $\mathcal{F}_0$. It holds by Theorem 16.2 for all nonnegative f that are measurable $\mathcal{F}_0$ because (Theorem 13.5) $0 \le f_n \uparrow f$ for simple functions f_n that are measurable $\mathcal{F}_0$. For functions measurable $\mathcal{F}_0$, integration with respect to μ is thus the same thing as integration with respect to μ_0. ■

In this example a property was extended by linearity from indicators to nonnegative simple functions and thence to the general nonnegative function by a monotone passage to the limit. This is a technique of very frequent application.

Example 16.5. For a finite or infinite sequence of measures μ_n on $\mathscr{F}$, $\mu(A) = \sum_n \mu_n(A)$ defines another measure (countably additive because [A27] sums can be reversed in a nonnegative double series). For indicators f,

$$\int f\,d\mu = \sum_n \int f\,d\mu_n,$$

and by linearity the same holds for simple $f \geq 0$. If $0 \leq f_k \uparrow f$ for simple f_k, then by Theorem 16.2 and Example 16.3, $\int f\,d\mu = \lim_k \int f_k\,d\mu = \lim_k \sum_n \int f_k\,d\mu_n = \sum_n \lim_k \int f_k\,d\mu_n = \sum_n \int f\,d\mu_n$. The relation in question thus holds for all nonnegative f. ■

An important consequence of the monotone convergence theorem is *Fatou's lemma*:

Theorem 16.3. *For nonnegative f_n,*

$$\int \liminf_n f_n\,d\mu \leq \liminf_n \int f_n\,d\mu. \tag{16.6}$$

PROOF. If $g_n = \inf_{k \geq n} f_k$, then $0 \leq g_n \uparrow g = \liminf_n f_n$, and the preceding two theorems give $\int f_n\,d\mu \geq \int g_n\,d\mu \to \int g\,d\mu$. ■

Example 16.6. On $(R^1, \mathscr{R}^1, \lambda)$, the functions $f_n = n^2 I_{(0, n^{-1})}$ and $f \equiv 0$ satisfy $f_n(x) \to f(x)$ for each x, but $\int f\,d\lambda = 0$ and $\int f_n\,d\lambda = n \to \infty$. This shows that the inequality in (16.6) can be strict and that it is not always possible to integrate to the limit. This phenomenon has been encountered before; see Examples 5.7 and 7.7. ■

Fatou's lemma leads to *Lebesgue's dominated convergence theorem*:

Theorem 16.4. *If $|f_n| \leq g$ almost everywhere, where g is integrable, and if $f_n \to f$ almost everywhere, then f and the f_n are integrable and $\int f_n\,d\mu \to \int f\,d\mu$.*

PROOF. Assume at the outset, not that the f_n converge, but only that they are dominated by an integrable g, which implies that all the f_n as well

as $f^* = \limsup_n f_n$ and $f_* = \liminf f_n$ are integrable. Since $g + f_n$ and $g - f_n$ are nonnegative, Fatou's lemma gives

$$\int g\,d\mu + \int f_*\,d\mu = \int \liminf_n (g + f_n)\,d\mu$$

$$\leq \liminf_n \int (g + f_n)\,d\mu = \int g\,d\mu + \liminf_n \int f_n\,d\mu,$$

and

$$\int g\,d\mu - \int f^*\,d\mu = \int \liminf_n (g - f_n)\,d\mu$$

$$\leq \liminf_n \int (g - f_n)\,d\mu = \int g\,d\mu - \limsup_n \int f_n\,d\mu.$$

Therefore

$$\int \liminf_n f_n\,d\mu \leq \liminf_n \int f_n\,d\mu \tag{16.7}$$

$$\leq \limsup_n \int f_n\,d\mu \leq \int \limsup_n f_n\,d\mu.$$

(Compare this with (4.9).)

Now use the assumption that $f_n \to f$ almost everywhere: f is dominated by g and hence is integrable, and the extreme terms in (16.7) agree with $\int f\,d\mu$. ■

Example 16.6 shows that this theorem can fail if no dominating g exists.

Example 16.7. *The Weierstrass M-test for series.* Consider the space $\{1, 2, \ldots\}$ together with counting measure, as in Example 16.1. If $|x_{nm}| \leq M_m$ and $\sum_m M_m < \infty$, and if $\lim_n x_{nm} = x_m$ for each m, then $\lim_n \sum_m x_{nm} = \sum_m x_m$. This follows by an application of Theorem 16.4 with the function given by the sequence $M_1, M_2, \ldots$ in the role of g. This is another standard result on infinite series [A28]. ■

The next result, the *bounded convergence theorem*, is a special case of Theorem 16.4. It contains Theorem 5.4 as a further special case.

Theorem 16.5. *If $\mu(\Omega) < \infty$ and the f_n are uniformly bounded, then $f_n \to f$ almost everywhere implies $\int f_n\,d\mu \to \int f\,d\mu$.*

The next two theorems are simply the series versions of the monotone and dominated convergence theorems.

Theorem 16.6. *If $f_n \geq 0$, then $\int \sum_n f_n \, d\mu = \sum_n \int f_n \, d\mu$.*

The members of this last equation are both equal either to ∞ or to the same finite, nonnegative real number.

Theorem 16.7. *If $\sum_n f_n$ converges almost everywhere and $|\sum_{k=1}^n f_k| \leq g$ almost everywhere, where g is integrable, then $\sum_n f_n$ and the f_n are integrable and $\int \sum_n f_n \, d\mu = \sum_n \int f_n \, d\mu$.*

Corollary. *If $\sum_n \int |f_n| \, d\mu < \infty$, then $\sum_n f_n$ converges absolutely almost everywhere and is integrable, and $\int \sum_n f_n \, d\mu = \sum_n \int f_n \, d\mu$.*

PROOF. The function $g = \sum_n |f_n|$ is integrable by Theorem 16.6 and is finite almost everywhere by Theorem 15.2(iii). Hence $\sum_n |f_n|$ and $\sum_n f_n$ converge almost everywhere, and Theorem 16.7 applies. ■

In place of a sequence $\{f_n\}$ of real measurable functions on $(\Omega, \mathcal{F}, \mu)$, consider a family $[f_t\colon t > 0]$ indexed by a continuous parameter t. Suppose of a measurable f that

$$\lim_{t \to \infty} f_t(\omega) = f(\omega) \tag{16.8}$$

on a set A, where

$$A \in \mathcal{F}, \qquad \mu(\Omega - A) = 0. \tag{16.9}$$

A technical point arises here, since $\mathcal{F}$ need not contain the ω-set where (16.8) holds:

Example 16.8. Let $\mathcal{F}$ consist of the Borel subsets of $\Omega = [0, 1)$, and let H be a nonmeasurable set—a subset of Ω that does not lie in $\mathcal{F}$ (see the end of Section 3). Define $f_t(\omega) = 1$ if ω equals the fractional part $t - \lfloor t \rfloor$ of t and their common value lies in H^c; define $f_t(\omega) = 0$ otherwise. Each f_t is measurable $\mathcal{F}$, but if $f(\omega) \equiv 0$, then the ω-set where (16.8) holds is exactly H. ■

Because of such examples, the set A above must be assumed to lie in $\mathcal{F}$. (Because of Theorem 13.4, no such assumption is necessary in the case of sequences.)

Suppose that f and the f_t are integrable. If $I_t = \int f_t \, d\mu$ converges to $I = \int f \, d\mu$ as $t \to \infty$, then certainly $I_{t_n} \to I$ for each sequence $\{t_n\}$ going to infinity. But the converse holds as well: If I_t does not converge to I, then there is a positive ϵ such that $|I_{t_n} - I| > \epsilon$ for a sequence $\{t_n\}$ going to infinity. To the question of whether I_{t_n} converges to I the previous theorems apply.

Suppose that (16.8) and $|f_t(\omega)| \leq g(\omega)$ both hold for $\omega \in A$, where A satisfies (16.9) and g is integrable. By the dominated convergence theorem, f and the f_t must then be integrable and $I_{t_n} \to I$ for each sequence $\{t_n\}$ going to infinity. It follows that $\int f_t \, d\mu \to \int f \, d\mu$. In this result t could go continuously to 0 or to some other value instead of to infinity.

Theorem 16.8. *Suppose that $f(\omega, t)$ is a measurable and integrable function of ω for each t in (a, b). Let $\phi(t) = \int f(\omega, t)\mu(d\omega)$.*

(i) *Suppose that for $\omega \in A$, where A satisfies (16.9), $f(\omega, t)$ is continuous in t at t_0; suppose further that $|f(\omega, t)| \le g(\omega)$ for $\omega \in A$ and $|t - t_0| < \delta$, where δ is independent of ω and g is integrable. Then $\phi(t)$ is continuous at t_0.*

(ii) *Suppose that for $\omega \in A$, where A satisfies (16.9), $f(\omega, t)$ has in (a, b) a derivative $f'(\omega, t)$; suppose further that $|f'(\omega, t)| \le g(\omega)$ for $\omega \in A$ and $t \in (a, b)$, where g is integrable. Then $\phi(t)$ has derivative $\int f'(\omega, t)\mu(d\omega)$ on (a, b).*

PROOF. Part (i) is an immediate consequence of the preceding discussion. To prove part (ii), consider a fixed t. If $\omega \in A$, then by the mean-value theorem,

$$\frac{f(\omega, t+h) - f(\omega, t)}{h} = f'(\omega, s),$$

where s lies between t and $t + h$. The ratio on the left goes[†] to $f'(\omega, t)$ as $h \to 0$ and is by hypothesis dominated by the integrable function $g(\omega)$. Therefore,

$$\frac{\phi(t+h) - \phi(t)}{h} = \int \frac{f(\omega, t+h) - f(\omega, t)}{h}\mu(d\omega) \to \int f'(\omega, t)\mu(d\omega). \qquad \blacksquare$$

The condition involving g in part (ii) can be weakened. It suffices to assume that for each t there is an integrable $g(\omega, t)$ such that $|f'(\omega, s)| \le g(\omega, t)$ for $\omega \in A$ and all s in some neighborhood of t.

Integration over Sets

The integral of f over a set A in $\mathscr{F}$ is defined by

$$\int_A f\,d\mu = \int I_A f\,d\mu. \tag{16.10}$$

The definition applies if f is defined only on A in the first place (set $f = 0$ outside A). Notice that $\int_A f\,d\mu = 0$ if $\mu(A) = 0$.

All the concepts and theorems above carry over in an obvious way to integrals over A. Theorems 16.6 and 16.7 yield this result:

Theorem 16.9. *If $A_1, A_2, \ldots$ are disjoint, and if f is either nonnegative or integrable, then $\int_{\cup_n A_n} f\,d\mu = \sum_n \int_{A_n} f\,d\mu$.*

[†]Letting h go to 0 through a sequence shows that each $f'(\cdot, t)$ is measurable $\mathscr{F}$ on A; take it to be 0, say, elsewhere.

The integrals (16.10) usually suffice to determine f:

Theorem 16.10. (i) *If f and g are nonnegative and $\int_A f\,d\mu = \int_A g\,d\mu$ for all A in $\mathscr{F}$, and if μ is σ-finite, then $f = g$ almost everywhere.*

(ii) *If f and g are integrable and $\int_A f\,d\mu = \int_A g\,d\mu$ for all A in $\mathscr{F}$, then $f = g$ almost everywhere.*

(iii) *If f and g are integrable and $\int_A f\,d\mu = \int_A g\,d\mu$ for all A in $\mathscr{P}$, where $\mathscr{P}$ is a π-system generating $\mathscr{F}$ and Ω is a finite or countable union of $\mathscr{P}$-sets, then $f = g$ almost everywhere.*

PROOF. Suppose that f and g are nonnegative and that $\int_A f\,d\mu \le \int_A g\,d\mu$ for all A in $\mathscr{F}$. If μ is σ-finite, there are $\mathscr{F}$-sets A_n such that $A_n \uparrow \Omega$ and $\mu(A_n) < \infty$. If $B_n = [0 \le g < f,\ g \le n]$, then the hypothesized inequality applied to $A_n \cap B_n$ implies $\int_{A_n \cap B_n} f\,d\mu \le \int_{A_n \cap B_n} g\,d\mu < \infty$ (finite because $A_n \cap B_n$ has finite measure and g is bounded there) and hence $\int I_{A_n \cap B_n}(f - g)\,d\mu = 0$. But then by Theorem 15.2(ii), the integrand is 0 almost everywhere, and so $\mu(A_n \cap B_n) = 0$. Therefore, $\mu[0 \le g < f,\ g < \infty] = 0$, so that $f \le g$ almost everywhere; (i) follows.

The argument for (ii) is simpler: If f and g are integrable and $\int_A f\,d\mu \le \int_A g\,d\mu$ for all A in $\mathscr{F}$, then $\int I_{[g<f]}(f - g)\,d\mu = 0$ and hence $\mu[g < f] = 0$ by Theorem 15.2(ii).

Part (iii) for nonnegative f and g follows from part (ii) together with Theorem 10.4. For the general case, prove that $f^+ + g^- = f^- + g^+$ almost everywhere. ■

Densities

Suppose that δ is a nonnegative measurable function and define a measure ν by (Theorem 16.9)

$$\nu(A) = \int_A \delta\,d\mu, \qquad A \in \mathscr{F}; \tag{16.11}$$

δ is not assumed integrable with respect to μ. Many measures arise in this way. Note that $\mu(A) = 0$ implies that $\nu(A) = 0$. Clearly, ν is finite if and only if δ is integrable μ. Another function δ' gives rise to the same ν if $\delta = \delta'$ almost everywhere. On the other hand, $\nu(A) = \int_A \delta'\,d\mu$ and (16.11) together imply that $\delta = \delta'$ almost everywhere if μ is σ-finite, as follows from Theorem 16.10(i).

The measure ν defined by (16.11) is said to have *density* δ with respect to μ. A density is by definition nonnegative.

Formal substitution $d\nu = \delta\,d\mu$ gives the formulas (16.12) and (16.13).

Theorem 16.11. *If ν has density δ with respect to μ, then*

$$\int f\,d\nu = \int f\delta\,d\mu \tag{16.12}$$

holds for nonnegative f. Moreover, f (not necessarily nonnegative) is integrable with respect to ν if and only if $f\delta$ is integrable with respect to μ, in which case (16.12) *and*

$$\int_A f\,d\nu = \int_A f\delta\,d\mu \tag{16.13}$$

both hold. For nonnegative f, (16.13) *always holds.*

Here $f\delta$ is to be taken as 0 if $f = 0$ or if $\delta = 0$; this is consistent with the conventions (15.2). Note that $\nu[\delta = 0] = 0$.

PROOF. If $f = I_A$, then $\int f\,d\nu = \nu(A)$, so that (16.12) reduces to the definition (16.11). If f is a simple nonnegative function, (16.12) then follows by linearity. If f is nonnegative, then $\int f_n\,d\nu = \int f_n\delta\,d\mu$ for the simple functions f_n of Theorem 13.5, and (16.12) follows by a monotone passage to the limit—that is, by Theorem 16.2. Note that both sides of (16.12) may be infinite.

Even if f is not nonnegative, (16.12) applies to $|f|$, whence it follows that f is integrable with respect to ν if and only if $f\delta$ is integrable with respect to μ. And if f is integrable, (16.12) follows from differencing the same result for f^+ and f^-. Replacing f by fI_A leads from (16.12) to (16.13). ■

Example 16.9. If $\nu(A) = \mu(A \cap A_0)$, then (16.11) holds with $\delta = I_{A_0}$, and (16.13) reduces to $\int_A f\,d\nu = \int_{A \cap A_0} f\,d\mu$. ■

Theorem 16.11 has two features in common with a number of theorems about integration:

(i) The relation in question, (16.12) in this case, in addition to holding for integrable functions, holds for all nonnegative functions—the point being that if one side of the equation is infinite, then so is the other, and if both are finite, then they have the same value. This is useful in checking for integrability in the first place.

(ii) The result is proved first for indicator functions, then for simple functions, then for nonnegative functions, then for integrable functions. In this connection, see Examples 16.4 and 16.5.

The next result is *Scheffé's theorem*.

Theorem 16.12. *Suppose that* $\nu_n(A) = \int_A \delta_n \, d\mu$ *and* $\nu(A) = \int_A \delta \, d\mu$ *for densities* δ_n *and* δ. *If*

$$\nu_n(\Omega) = \nu(\Omega) < \infty, \qquad n = 1, 2, \ldots, \tag{16.14}$$

and if $\delta_n \to \delta$ *except on a set of* μ*-measure* 0, *then*

$$\sup_{A \in \mathcal{F}} |\nu(A) - \nu_n(A)| \leq \int_\Omega |\delta - \delta_n| \, d\mu \to 0. \tag{16.15}$$

PROOF. The inequality in (16.15) of course follows from (16.5). Let $g_n = \delta - \delta_n$. The positive part g_n^+ of g_n converges to 0 except on a set of μ-measure 0. Moreover, $0 \leq g_n^+ \leq \delta$ and δ is integrable, and so the dominated convergence theorem applies: $\int g_n^+ \, d\mu \to 0$. But $\int g_n \, d\mu = 0$ by (16.14), and therefore

$$\begin{aligned} \int_\Omega |g_n| \, d\mu &= \int_{[g_n \geq 0]} g_n \, d\mu - \int_{[g_n < 0]} g_n \, d\mu \\ &= 2 \int_{[g_n \geq 0]} g_n \, d\mu = 2 \int_\Omega g_n^+ \, d\mu \to 0. \end{aligned}$$

■

A corollary concerning infinite series follows immediately—take μ as counting measure on $\Omega = \{1, 2, \ldots\}$.

Corollary. *If* $\sum_m x_{nm} = \sum_m x_m < \infty$, *the terms being nonnegative, and if* $\lim_n x_{nm} = x_m$ *for each* m, *then* $\lim_n \sum_m |x_{nm} - x_m| = 0$. *If* y_m *is bounded, then* $\lim_n \sum_m y_m x_{nm} = \sum_m y_m x_m$.

Change of Variable

Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces, and suppose that the mapping $T: \Omega \to \Omega'$ is measurable $\mathcal{F}/\mathcal{F}'$. For a measure μ on $\mathcal{F}$, define a measure μT^{-1} on $\mathcal{F}'$ by

$$\mu T^{-1}(A') = \mu(T^{-1}A'), \qquad A' \in \mathcal{F}', \tag{16.16}$$

as at the end of Section 13.

Suppose f is a real function on Ω' that is measurable $\mathcal{F}'$, so that the composition fT is a real function on Ω that is measurable $\mathcal{F}$ (Theorem 13.1(ii)). The change-of-variable formulas are (16.17) and (16.18). If $A' = \Omega'$, the second reduces to the first.

Theorem 16.13. *If f is nonnegative, then*

$$\int_{\Omega} f(T\omega)\mu(d\omega) = \int_{\Omega'} f(\omega')\mu T^{-1}(d\omega'). \tag{16.17}$$

A function f (not necessarily nonnegative) is integrable with respect to μT^{-1} if and only if fT is integrable with respect to μ, in which case (16.17) *and*

$$\int_{T^{-1}A'} f(T\omega)\mu(d\omega) = \int_{A'} f(\omega')\mu T^{-1}(d\omega') \tag{16.18}$$

hold. For nonnegative f, (16.18) *always holds.*

PROOF. If $f = I_{A'}$, then $fT = I_{T^{-1}A'}$, and so (16.17) reduces to the definition (16.16). By linearity, (16.17) holds for nonnegative simple functions. If f_n are simple functions for which $0 \le f_n \uparrow f$, then $0 \le f_n T \uparrow fT$, and (16.17) follows by the monotone convergence theorem.

An application of (16.17) to $|f|$ establishes the assertion about integrability, and for integrable f, (16.17) follows by decomposition into positive and negative parts. Finally, if f is replaced by $fI_{A'}$, (16.17) reduces to (16.18). ■

Example 16.10. Suppose that $(\Omega', \mathscr{F}') = (R^1, \mathscr{R}^1)$ and $T = \varphi$ is an ordinary real function, measurable $\mathscr{F}$. If $f(x) = x$, (16.17) becomes

$$\int_{\Omega} \varphi(\omega)\mu(d\omega) = \int_{R^1} x\mu\varphi^{-1}(dx). \tag{16.19}$$

If $\varphi = \sum_i x_i I_{A_i}$ is simple, then $\mu\varphi^{-1}$ has mass $\mu(A_i)$ at x_i, and each side of (16.19) reduces to $\sum_i x_i \mu(A_i)$. ■

Uniform Integrability

If f is integrable, then $|f|I_{[|f|\ge\alpha]}$ goes to 0 almost everywhere as $\alpha \to \infty$ and is dominated by $|f|$, and hence

$$\lim_{\alpha\to\infty} \int_{[|f|\ge\alpha]} |f|\, d\mu = 0. \tag{16.20}$$

A sequence $\{f_n\}$ is *uniformly integrable* if (16.20) holds uniformly in n:

$$\lim_{\alpha\to\infty} \sup_n \int_{[|f_n|\ge\alpha]} |f_n|\, d\mu = 0. \tag{16.21}$$

If (16.21) holds and $\mu(\Omega) < \infty$, and if α is large enough that the supremum in (16.21) is less than 1, then

$$\int |f_n|\, d\mu \le \alpha\mu(\Omega) + 1, \tag{16.22}$$

and hence the f_n are integrable. On the other hand, (16.21) always holds if the f_n are uniformly bounded, but the f_n need not in that case be integrable if $\mu(\Omega) = \infty$. For this reason the concept of uniform integrability is interesting only for μ finite.

If h is the maximum of $|f|$ and $|g|$, then

$$\int_{|f+g|\ge 2\alpha} |f+g|\, d\mu \le 2\int_{h\ge\alpha} h\, d\mu \le 2\int_{|f|\ge\alpha} |f|\, d\mu + 2\int_{|g|\ge\alpha} |g|\, d\mu.$$

Therefore, if $\{f_n\}$ and $\{g_n\}$ are uniformly integrable, so is $\{f_n + g_n\}$.

Theorem 16.14. *Suppose that $\mu(\Omega) < \infty$ and $f_n \to f$ almost everywhere.*

(i) *If the f_n are uniformly integrable, then f is integrable and*

$$\int f_n\, d\mu \to \int f d\mu. \tag{16.23}$$

(ii) *If f and the f_n are nonnegative and integrable, then* (16.23) *implies that the f_n are uniformly integrable.*

PROOF. If the f_n are uniformly integrable, it follows by (16.22) and Fatou's lemma that f is integrable. Define

$$f_n^{(\alpha)} = \begin{cases} f_n & \text{if } |f_n| < \alpha, \\ 0 & \text{if } |f_n| \ge \alpha, \end{cases} \qquad f^{(\alpha)} = \begin{cases} f & \text{if } |f| < \alpha, \\ 0 & \text{if } |f| \ge \alpha. \end{cases}$$

If $\mu[|f| = \alpha] = 0$, then $f_n^{(\alpha)} \to f^{(\alpha)}$ almost everywhere, and by the bounded convergence theorem,

$$\int f_n^{(\alpha)}\, d\mu \to \int f^{(\alpha)}\, d\mu. \tag{16.24}$$

Since

$$\int f_n\, d\mu - \int f_n^{(\alpha)}\, d\mu = \int_{|f_n|\ge\alpha} f_n\, d\mu \tag{16.25}$$

and

$$\int f d\mu - \int f^{(\alpha)}\, d\mu = \int_{|f|\ge\alpha} f d\mu, \tag{16.26}$$

it follows from (16.24) that

$$\limsup_n \left| \int f_n \, d\mu - \int f \, d\mu \right| \leq \sup_n \int_{|f_n| \geq \alpha} |f_n| \, d\mu + \int_{|f| \geq \alpha} |f| \, d\mu .$$

And now (16.23) follows from the uniform integrability and the fact that $\mu[|f| = \alpha] = 0$ for all but countably many α.

Suppose on the other hand that (16.23) holds, where f and the f_n are nonnegative and integrable. If $\mu[f = \alpha] = 0$, then (16.24) holds, and from (16.25) and (16.26) follows

$$\int_{f_n \geq \alpha} f_n \, d\mu \to \int_{f \geq \alpha} f \, d\mu . \tag{16.27}$$

Since f is integrable, there is, for given ϵ, an α such that the limit in (16.27) is less than ϵ and $\mu[f = \alpha] = 0$. But then the integral on the left is less than ϵ for all n exceeding some n_0. Since the f_n are individually integrable, uniform integrability follows (increase α). ■

Corollary. *Suppose that $\mu(\Omega) < \infty$. If f and the f_n are integrable, and if $f_n \to f$ almost everywhere, then these conditions are equivalent:*

(i) *f_n are uniformly integrable;*
(ii) $\int |f - f_n| \, d\mu \to 0$;
(iii) $\int |f_n| \, d\mu \to \int |f| \, d\mu$.

PROOF. If (i) holds, then the differences $|f - f_n|$ are uniformly integrable, and since they converge to 0 almost everywhere, (ii) follows by the theorem. And (ii) implies (iii) because $||f| - |f_n|| \leq |f - f_n|$. Finally, it follows from the theorem that (iii) implies (i). ■

Suppose that

$$\sup_n \int |f_n|^{1+\epsilon} \, d\mu < \infty \tag{16.28}$$

for a positive ϵ. If K is the supremum here, then

$$\int_{[|f_n| \geq \alpha]} |f_n| \, d\mu \leq \frac{1}{\alpha^\epsilon} \int_{[|f_n| \geq \alpha]} |f_n|^{1+\epsilon} \, d\mu \leq \frac{K}{\alpha^\epsilon} ,$$

and so $\{f_n\}$ is uniformly integrable.

Complex Functions

A complex-valued function on Ω has the form $f(\omega) = g(\omega) + ih(\omega)$, where g and h are ordinary finite-valued real functions on Ω. It is, by definition, measurable $\mathscr{F}$ if g and h are. If g and h are integrable, then f is by definition integrable, and its integral is of course taken as

$$\int (g + ih) \, d\mu = \int g \, d\mu + i \int h \, d\mu . \tag{16.29}$$

Since $\max\{|g|, |h|\} \le |f| \le |g| + |h|$, f is integrable if and only if $\int |f|\, d\mu < \infty$, just as in the real case.

The linearity equation (16.3) extends to complex functions and coefficients—the proof requires only that everything be decomposed into real and imaginary parts. Consider the inequality (16.4) for the complex case:

$$\left| \int f\, d\mu \right| \le \int |f|\, d\mu. \tag{16.30}$$

If $f = g + ih$ and g and h are simple, the corresponding partitions can be taken to be the same ($g = \sum_k x_k I_{A_k}$ and $h = \sum_k y_k I_{A_k}$), and (16.30) follows by the triangle inequality. For the general integrable f, represent g and h as limits of simple functions dominated by $|f|$, and pass to the limit.

The results on integration to the limit extend as well. Suppose that $f_k = g_k + ih_k$ are complex functions satisfying $\sum_k \int |f_k|\, d\mu < \infty$. Then $\sum_k \int |g_k|\, d\mu < \infty$, and so by the corollary to Theorem 16.7, $\sum_k g_k$ is integrable and integrates to $\sum_k \int g_k\, d\mu$. The same is true of the imaginary parts, and hence $\sum_k f_k$ is integrable and

$$\int \sum_k f_k\, d\mu = \sum_k \int f_k\, d\mu. \tag{16.31}$$

PROBLEMS

16.1. 13.9↑ Suppose that $\mu(\Omega) < \infty$ and f_n are uniformly bounded.

(a) Assume $f_n \to f$ uniformly and deduce $\int f_n\, d\mu \to \int f\, d\mu$ from (16.5).

(b) Use part (a) and Egoroff's theorem to give another proof of Theorem 16.5.

16.2. Prove that if $0 \le f_n \to f$ almost everywhere and $\int f_n\, d\mu \le A < \infty$, then f is integrable and $\int f\, d\mu \le A$. (This is essentially the same as Fatou's lemma and is sometimes called by that name.)

16.3. Suppose that f_n are integrable and $\sup_n \int f_n\, d\mu < \infty$. Show that, if $f_n \uparrow f$, then f is integrable and $\int f_n\, d\mu \to \int f\, d\mu$. This is *Beppo Levi's theorem*.

16.4. (a) Suppose that functions a_n, b_n, f_n converge almost everywhere to functions a, b, f, respectively. Suppose that the first two sequences may be integrated to the limit—that is, the functions are all integrable and $\int a_n\, d\mu \to \int a\, d\mu$, $\int b_n\, d\mu \to \int b\, d\mu$. Suppose, finally, that the first two sequences enclose the third: $a_n \le f_n \le b_n$ almost everywhere. Show that the third may be integrated to the limit.

(b) Deduce Lebesgue's dominated convergence theorem from part (a).

16.5. About Theorem 16.8:

(a) Part (i) is local: there can be a different set A for each t_0. Part (ii) can be recast as a local theorem. Suppose that for $\omega \in A$, where A satisfies (16.9),

$f(\omega, t)$ has derivative $f'(\omega, t_0)$ at t_0; suppose further that

$$\left|\frac{f(\omega, t_0+h)-f(\omega, t_0)}{h}\right| \le g_1(\omega) \tag{16.32}$$

for $\omega \in A$ and $0<|h|<\delta$, where δ is independent of ω and g_1 is integrable. Then $\phi'(t_0)=\int f'(\omega, t_0)\mu(d\omega)$.

The natural way to check (16.32), however, is by the mean-value theorem, and this requires (for $\omega \in A$) a derivative throughout a neighborhood of t_0.

(b) If μ is Lebesgue measure on the unit interval Ω, $(a, b)=(0,1)$, and $f(\omega, t)=I_{(0,t)}(\omega)$, then part (i) applies but part (ii) does not. Why? What about (16.32)?

16.6. Suppose that $f(\omega, \cdot)$ is, for each ω, a function on an open set W in the complex plane and that $f(\cdot, z)$ is for z in W measurable $\mathscr{F}$ and integrable. Suppose that A satisfies (16.9), that $f(\omega, \cdot)$ is analytic in W for ω in A, and that for each z_0 in W there is an integrable $g(\cdot, z_0)$ such that $|f'(\omega, z)| \le g(\omega, z_0)$ for all $\omega \in A$ and all z in some neighborhood of z_0. Show that $\int f(\omega, z)\mu(d\omega)$ is analytic in W.

16.7. (a) Show that if $|f_n| \le g$ and g is integrable, then $\{f_n\}$ is uniformly integrable. Compare the hypotheses of Theorems 16.4 and 16.14.

(b) On the unit interval with Lebesgue measure, let $f_n=(n/\log n)I_{(0,n^{-1})}$ for $n \ge 3$. Show that the f_n are uniformly integrable (and $\int f_n\, d\mu \to 0$) although they are not dominated by any integrable g.

(c) Show for $f_n=nI_{(0,n^{-1}0)}-nI_{(n^{-1},2n^{-1})}$ that the f_n can be integrated to the limit (Lebesgue measure) even though the f_n are not uniformly integrable.

16.8. Show that if f is integrable, then for each ϵ there is a δ such that $\mu(A)<\delta$ implies $\int_A |f|\, d\mu < \epsilon$.

16.9. ↑ Suppose that $\mu(\Omega)<\infty$. Show that $\{f_n\}$ is uniformly integrable if and only if (i) $\int |f_n|\, d\mu$ is bounded and (ii) for each ϵ there is a δ such that $\mu(A)<\delta$ implies $\int_A |f_n|\, d\mu < \epsilon$ for all n.

16.10. 2.19 16.9 ↑ Assume $\mu(\Omega)<\infty$.

(a) Show by examples that neither of the conditions (i) and (ii) in the preceding problem implies the other.

(b) Show that (ii) implies (i) for all sequences $\{f_n\}$ if and only if μ is nonatomic.

16.11. Let f be a complex-valued function integrating to $re^{i\theta}$, $r \ge 0$. From $\int(|f(\omega)|-e^{-i\theta}f(\omega))\mu(d\mu)=\int |f|\, d\mu - r$, deduce (16.30).

16.12. 11.5 ↑ Consider the vector lattice $\mathscr{L}$ and the functional Λ of Problems 11.4 and 11.5. Let μ be the extension (Theorem 11.3) to $\mathscr{F}=\sigma(\mathscr{F}_0)$ of the set function μ_0 on $\mathscr{F}_0$.

(a) Show by (11.7) that for positive x, y_1, y_2 one has $\nu([f>1]\times(0,x])=x\mu_0[f>1]=x\mu[f>1]$ and $\nu([y_1<f\le y_2]\times(0,x])=x\mu[y_1<f\le y_2]$.

(b) Show that if $f \in \mathscr{L}$, then f is integrable and

$$\Lambda(f) = \int f \, d\mu.$$

(Consider first the case $f \geq 0$.) This is the *Daniell–Stone representation theorem*.

SECTION 17. THE INTEGRAL WITH RESPECT TO LEBESGUE MEASURE

The Lebesgue Integral on the Line

A real measurable function on the line is *Lebesgue integrable* if it is integrable with respect to Lebesgue measure λ, and its *Lebesgue integral* $\int f \, d\lambda$ is denoted by $\int f(x)\, dx$, or, in the case of integration over an interval, by $\int_a^b f(x)\, dx$. The theory of the preceding two sections of course applies to the Lebesgue integral. It is instructive to compare it with the Riemann integral.

The Riemann Integral

A real function f on an interval $(a, b]$ is by definition[†] *Riemann integrable*, with integral r, if this condition holds: For each ϵ there exists a δ with the property that

$$\left| r - \sum_i f(x_i)\lambda(I_i) \right| < \epsilon \tag{17.1}$$

if $\{I_i\}$ is any finite partition of $(a, b]$ into subintervals satisfying $\lambda(I_i) < \delta$ and if $x_i \in I_i$ for each i. The Riemann integral for step functions was used in Section 1.

Suppose that f is Borel measurable, and suppose that f is bounded, so that it is Lebesgue integrable. If f is also Riemann integrable, the r of (17.1) must coincide with the Lebesgue integral $\int_a^b f(x)\, dx$. To see this, first note that letting x_i vary over I_i leads from (17.1) to

$$\left| r - \sum_i \sup_{x \in I_i} f(x) \cdot \lambda(I_i) \right| \leq \epsilon. \tag{17.2}$$

Consider the simple function g with value $\sup_{x \in I_i} f(x)$ on I_i. Now $f \leq g$, and the sum in (17.2) is the Lebesgue integral of g. By monotonicity of the

[†]For other definitions, see the first problem at the end of the section and the *Note on terminology* following it.

Lebesgue integral, $\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx \le r + \epsilon$. The reverse inequality follows in the same way, and so $\int_a^b f(x)\,dx = r$. *Therefore, the Riemann integral when it exists coincides with the Lebsegue integral.*

Suppose that f is continuous on $[a, b]$. By uniform continuity, for each ϵ there exists a δ such that $|f(x) - f(y)| < \epsilon/(b-a)$ if $|x - y| < \delta$. If $\lambda(I_i) < \delta$ and $x_i \in I_i$, then $g = \sum_i f(x_i) I_{I_i}$ satisfies $|f - g| < \epsilon/(b-a)$ and hence $|\int_a^b f\,dx - \int_a^b g\,dx| < \epsilon$. But this is (17.1) with r replaced (as it must be) by the Lebesgue integral $\int_a^b f\,dx$: *A continuous function on a closed interval is Riemann integrable.*

Example 17.1. If f is the indicator of the set of rationals in $(0, 1]$, then the Lebesgue integral $\int_0^1 (x)\,dx$ is 0 because $f = 0$ almost everywhere. But for an arbitrary partition $\{I_i\}$ of $(0, 1]$ into intervals, $\sum_i f(x_i)\lambda(I_i)$ with $x_i \in I_i$ is 1 if each x_i is taken from the rationals and 0 if each x_i is taken from the irrationals. Thus f is not Riemann integrable. ■

Example 17.2. For the f of Example 17.1, there exists a g (namely, $g \equiv 0$) such that $f = g$ almost everywhere and g is Riemann integrable. To show that the Lebesgue theory is not reducible to the Riemann theory by the casting out of sets of measure 0, it is of interest to produce an f (bounded and Borel measurable) for which no such g exists.

In Examples 3.1 and 3.2 there were constructed Borel subsets A of $(0, 1]$ such that $0 < \lambda(A) < 1$ and such that $\lambda(A \cap I) > 0$ for each subinterval I of $(0, 1]$. Take $f = I_A$. Suppose that $f = g$ almost everywhere and that $\{I_i\}$ is a decomposition of $(0, 1]$ into subintervals. Since $\lambda(I_i \cap A \cap [f = g]) = \lambda(I_i \cap A) > 0$, it follows that $g(y_i) = f(y_i) = 1$ for some y_i in $I_i \cap A$, and therefore,

$$\sum_i g(y_i)\lambda(I_i) = 1 > \lambda(A). \tag{17.3}$$

If g were Riemann integrable, its Riemann integral would coincide with the Lebesgue integrals $\int g\,dx = \int f\,dx = \lambda(A)$, in contradiction to (17.3). ■

It is because of their extreme oscillations that the functions in Examples 17.1 and 17.2 fail to be Riemann integrable. (It can be shown that a bounded function on a bounded interval is Riemann integrable if and only if the set of its discontinuities has Lebesgue measure 0.†) This cannot happen in the case of the Lebesgue integral of a measurable function: if f fails to be Lebesgue integrable, it is because its positive part or its negative part is too large, not because one or the other is too irregular.

†See Problem 17.1.

Example 17.3. It is an important analytic fact that

$$\lim_{t\to\infty}\int_0^t \frac{\sin x}{x}\,dx = \frac{\pi}{2}. \tag{17.4}$$

The existence of the limit is simple to prove, because $\int_{(n-1)\pi}^{n\pi} x^{-1}\sin x\,dx$ alternates in sign and its absolute value decreases to 0; the value of the limit will be identified in the next section (Example 18.4). On the other hand, $x^{-1}\sin x$ is not Lebesgue integrable over $(0,\infty)$, because its positive and negative parts integrate to ∞. Within the conventions of the Lebesgue theory, (17.4) thus cannot be written $\int_0^\infty x^{-1}\sin x\,dx = \pi/2$—although such "improper" integrals appear in calculus texts. It is, of course, just a question of choosing the terminology most convenient for the subject at hand. ■

The function in Example 17.2 is not equal almost everywhere to any Riemann integrable function. Every Lebesgue integrable function can, however, be approximated in a certain sense by Riemann integrable functions of two kinds.

Theorem 17.1. *Suppose that* $\int|f|\,dx < \infty$ *and* $\epsilon > 0$.

(i) *There is a step function* $g = \sum_{i=1}^k x_i I_{A_i}$, *with bounded intervals as the* A_i, *such that* $\int|f-g|\,dx < \epsilon$.

(ii) *There is a continuous integrable* h *with bounded support such that* $\int|f-h|\,dx < \epsilon$.

PROOF. By the construction (13.6) and the dominated convergence theorem, (i) holds if the A_i are not required to be intervals; moreover, $\lambda(A_i) < \infty$ for each i for which $x_i \neq 0$. By Theorem 11.4 there exists a finite disjoint union B_i of intervals such that $\lambda(A_i \Delta B_i) < \epsilon/k|x_i|$. But then $\sum_i x_i I_{B_i}$ satisfies the requirements of (i) with 2ϵ in place of ϵ.

To prove (ii) it is only necessary to show that for the g of (i) there is a continuous h such that $\int|g-h|\,dx < \epsilon$. Suppose that $A_i = (a_i, b_i]$; let $h_i(x)$ be 1 on $(a_i, b_i]$ and 0 outside $(a_i-\delta, b_i+\delta]$, and let it increase linearly from 0 to 1 over $(a_i-\delta, a_i]$ and decrease linearly from 1 to 0 over $(b_i, b_i+\delta]$. Since $\int|I_{A_i}-h_i|\,dx \to 0$ as $\delta\to 0$, $h = \sum x_i h_i$ for sufficiently small δ will satisfy the requirements. ■

The Lebesgue integral is thus determined by its values for continuous functions.†

†This provides another way of *defining* the Lebesgue integral on the line. See Problem 17.13.

The Fundamental Theorem of Calculus

Adopt the convention that $\int_\alpha^\beta = -\int_\beta^\alpha$ if $\alpha > \beta$. For positive h,

$$\left|\frac{1}{h}\int_x^{x+h} f(y)\,dy - f(x)\right| \le \frac{1}{h}\int_x^{x+h} |f(y) - f(x)|\,dy$$
$$\le \sup[|f(y) - f(x)|: x \le y \le x+h],$$

and the right side goes to 0 with h if f is continuous at x. The same thing holds for negative h, and therefore $\int_a^x f(y)\,dy$ has derivative $f(x)$:

$$\frac{d}{dx}\int_a^x f(y)\,dy = f(x) \tag{17.5}$$

if f is continuous at x.

Suppose that F is a function with continuous derivative $F' = f$; that is, suppose that F is a *primitive* of the continuous function f. Then

$$\int_a^b f(x)\,dx = \int_a^b F'(x)\,dx = F(b) - F(a), \tag{17.6}$$

as follows from the fact that $F(x) - F(a)$ and $\int_a^x f(y)\,dy$ agree at $x = a$ and by (17.5) have identical derivatives. For continuous f, (17.5) and (17.6) are two ways of stating the fundamental theorem of calculus. To the calculation of Lebesgue integrals the methods of elementary calculus thus apply.

As will follow from the general theory of derivatives in Section 31, (17.5) holds outside a set of Lebesgue measure 0 if f is integrable—it need not be continuous. As the following example shows, however, (17.6) can fail for discontinuous f.

Example 17.4. Define $F(x) = x^2 \sin x^{-2}$ for $0 \le x \le \frac{1}{2}$ and $F(x) = 0$ for $x \le 0$ and for $x \ge 1$. Now for $\frac{1}{2} < x < 1$ define $F(x)$ in such a way that F is continuously differentiable over $(0, \infty)$. Then F is everywhere differentiable, but $F'(0) = 0$ and $F'(x) = 2x \sin x^{-2} - 2x^{-1} \cos x^{-2}$ for $0 < x < \frac{1}{2}$. Thus F' is discontinuous at 0; F' is, in fact, not even integrable over $(0, 1]$, which makes (17.6) impossible for $a = 0$.

For a more extreme example, decompose $(0, 1]$ into countably many subintervals $(a_n, b_n]$. Define $G(x) = 0$ for $x \le 0$ and $x \ge 1$, and on $(a_n, b_n]$ define $G(x) = F((x - a_n)/(b_n - a_n))$. Then G is everywhere differentiable, but (17.6) is impossible for G if $(a, b]$ contains any of the $(a_n, b_n]$, because G is not integrable over any of them. ■

Change of Variable

For

$$[a, b] \xrightarrow{T} [u, v] \xrightarrow{f} R^1, \tag{17.7}$$

the change-of-variable formula is

$$\int_a^b f(Tx)T'(x)\,dx = \int_{Ta}^{Tb} f(y)\,dy. \tag{17.8}$$

If T' exists and is continuous, and if f is continuous, the two integrals are finite because the integrands are bounded, and to prove (17.8) it is enough to let b be a variable and differentiate with respect to it.†

With the obvious limiting arguments, this applies to unbounded intervals and to open ones:

Example 17.5. Put $T(x) = \tan x$ on $(-\pi/2, \pi/2)$. Then $T'(x) = 1 + T^2(x)$, and (17.8) for $f(y) = (1+y^2)^{-1}$ gives

$$\int_{-\infty}^{\infty} \frac{dy}{1+y^2} = \pi. \tag{17.9}$$

■

The Lebesgue Integral in R^k

The k-dimensional Lebesgue integral, the integral in $(R^k, \mathscr{R}^k, \lambda_k)$, is denoted $\int f(x)\,dx$, it being understood that $x = (x_1, \ldots, x_k)$. In low-dimensional cases it is also denoted $\int\int_A f(x_1, x_2)\,dx_1\,dx_2$, and so on.

As for the rule for changing variables, suppose that $T\colon U \to R^k$, where U is an open set in R^k. The map has the form $Tx = (t_1(x), \ldots, t_k(x))$; it is by definition continuously differentiable if the partial derivatives $t_{ij}(x) = \partial t_i/\partial x_j$ exist and are continuous in U. Let $D_x = [t_{ij}(x)]$ be the Jacobian matrix, let $J(x) = \det D_x$ be the Jacobian determinant, and let $V = TU$.

Theorem 17.2. *Let T be a continuously differentiable map of the open set U onto V. Suppose that T is one-to-one and that $J(x) \neq 0$ for all x. If f is nonnegative, then*

$$\int_U f(Tx)|J(x)|\,dx = \int_V f(y)\,dy. \tag{17.10}$$

By the inverse-function theorem [A35], V is open and the inverse point mapping T^{-1} is continuously differentiable. It is assumed in (17.10) that $f\colon V \to R^1$ is a Borel function. As usual, for the general f, (17.10) holds with $|f|$ in place of f, and if the two sides are finite, the absolute-value bars can be removed; and of course f can be replaced by fI_B or fI_{TA}.

†See Problem 17.11 for extensions.

Example 17.6. Suppose that T is a nonsingular linear transformation on $U = V = R^k$. Then D_x is for each x the matrix of the transformation. If T is identified with this matrix, then (17.10) becomes

$$|\det T| \int_U f(Tx)\,dx = \int_V f(y)\,dy. \tag{17.11}$$

If $f = I_{TA}$, this holds because of (12.2), and then it follows in the usual sequence for simple f and for the general nonnegative f: Theorem 17.2 is easy in the linear case. ■

Example 17.7. In R^2 take $U = [(\rho, \theta)\colon \rho > 0,\ 0 < \theta < 2\pi]$ and $T(\rho, \theta) = (\rho \cos\theta, \rho \sin\theta)$. The Jacobian is $J(\rho, \theta) = \rho$, and (17.10) gives the formula for integrating in polar coordinates:

$$\int\int_{\substack{\rho>0 \\ 0<\theta<2\pi}} f(\rho\cos\theta, \rho\sin\theta)\,\rho\,d\rho\,d\theta = \int\int_{R^2} f(x,y)\,dx\,dy. \tag{17.12}$$

Here V is R^2 with the ray $[(x,0)\colon x \geq 0]$ removed; (17.12) obviously holds even though the ray is included on the right. If the constraint on θ is replaced by $0 < \theta < 4\pi$, for example, then (17.12) is false (a factor of 2 is needed on the right). This explains the assumption that T is one-to-one. ■

Theorem 17.2 is not the strongest possible; it is only necessary to assume that T is one-to-one on the set $U_0 = [x \in U\colon J(x) \neq 0]$. This is because, by Sard's theorem,[†] $\lambda_k(V - TU_0) = 0$.

PROOF OF THEOREM 17.2. Suppose it is shown that

$$\int_U f(Tx)|J(x)|\,dx \geq \int_V f(y)\,dy \tag{17.13}$$

for nonnegative f. Apply this to $T^{-1}\colon V \to U$, which [A35] is continuously differentiable and has Jacobian $J^{-1}(T^{-1}y)$ at y:

$$\int_V g(T^{-1}y)|J^{-1}(T^{-1}y)|\,dy \geq \int_U g(x)\,dx$$

for nonnegative g on V. Taking $g(x) = f(Tx)|J(x)|$ here leads back to (17.13), but with the inequality reversed. Therefore, proving (17.13) will be enough.

For $f = I_{TA}$, (17.13) reduces to

$$\int_A |J(x)|\,dx \geq \lambda_k(TA). \tag{17.14}$$

[†]SPIVAK, p. 72.

Each side of (17.14) is a measure on $\mathscr{U} = U \cap \mathscr{R}^k$. If $\mathscr{A}$ consists of the rectangles A satisfying $A^- \subset U$, then $\mathscr{A}$ is a semiring generating $\mathscr{U}$, U is a countable union of $\mathscr{A}$-sets, and the left side of (17.14) is finite for A in $\mathscr{A}$ ($\sup_A |J| < \infty$). It follows by Corollary 2 to Theorem 11.4 that if (17.14) holds for A in $\mathscr{A}$, then it holds for A in $\mathscr{U}$. But then (linearity and monotone convergence) (17.13) will follow.

Proof of (17.14) *for A in* $\mathscr{A}$. Split the given rectangle A into finitely many subrectangles Q_i satisfying

$$\operatorname{diam} Q_i < \delta, \tag{17.15}$$

δ to be determined. Let x_i be some point of Q_i. Given ϵ, choose δ in the first place so that $|J(x) - J(x')| < \epsilon$ if $x, x' \in A^-$ and $|x - x'| < \delta$. Then (17.15) implies

$$\sum_i |J(x_i)| \lambda_k(Q_i) \le \int_A |J(x)|\, dx + \epsilon \lambda_k(A). \tag{17.16}$$

Let $Q_i^{+\epsilon}$ be a rectangle that is concentric with Q_i and similar to it and whose edge lengths are those of Q_i multiplied by $1 + \epsilon$. For x in U consider the affine transformation

$$\phi_x z = D_x(z - x) + Tx, \qquad z \in R^k; \tag{17.17}$$

$\phi_x z$ will [A34] be a good approximation to Tz for z near x. Suppose, as will be proved in a moment, that for each ϵ there is a δ such that, if (17.15) holds, then, for each i, ϕ_{x_i} approximates T so well on Q_i that

$$TQ_i \subset \phi_{x_i} Q_i^{+\epsilon}. \tag{17.18}$$

By Theorem 12.2, which shows in the nonsingular case how an affine transformation changes the Lebesgue measures of sets, $\lambda_k(\phi_{x_i} Q_i^{+\epsilon}) = |J(x_i)| \lambda_k(Q_i^{+\epsilon})$. If (17.18) holds, then

$$\begin{aligned} \lambda_k(TA) &= \sum_i \lambda_k(TQ_i) \le \sum_i \lambda_k\bigl(\phi_{x_i} Q_i^{+\epsilon}\bigr) \\ &= \sum_i |J(x_i)| \lambda_k(Q_i^{+\epsilon}) = (1+\epsilon)^k \sum_i |J(x_i)| \lambda_k(Q_i). \end{aligned} \tag{17.19}$$

(This, the central step in the proof, shows where the Jacobian in (17.10) comes from.) If for each ϵ there is a δ such that (17.15) implies both (17.16) and (17.19), then (17.14) will follow. Thus everything depends on (17.18), and the remaining problem is to show that for each ϵ there is a δ such that (17.18) holds if (17.15) does.

Proof of (17.18). As (x, z) varies over the compact set $A^- \times [z\colon |z| = 1]$, $|D_x^{-1} z|$ is continuous, and therefore, for some c,

$$|D_x^{-1} z| \le c|z| \qquad \text{for } x \in A,\ z \in R^k. \tag{17.20}$$

Since the t_{jl} are uniformly continuous on A^-, δ can be chosen so that $|t_{jl}(z) - t_{jl}(x)| \le \epsilon / k^2 c$ for all j, l if $z, x \in A$ and $|z - x| < \delta$. But then, by linear approximation [A34: (16)], $|Tz - Tx - D_x(z - x)| \le \epsilon c^{-1} |z - x| < \epsilon c^{-1} \delta$. If (17.15) holds and $\delta < 1$, then by the definition (17.17),

$$|Tz - \phi_{x_i} z| < \epsilon / c \qquad \text{for } z \in Q_i. \tag{17.21}$$

To prove (17.18), note that $z \in Q_i$ implies

$$|\phi_{x_i}^{-1}Tz - z| = |\phi_{x_i}^{-1}Tz - \phi_{x_i}^{-1}\phi_{x_i}z| = |D_{x_i}^{-1}(Tz - \phi_{x_i}z)|$$
$$\leq c|Tz - \phi_{x_i}z| < \epsilon,$$

where the first inequality follows by (17.20) and the second by (17.21). Since $\phi_{x_i}^{-1}Tz$ is within ϵ of the point z of Q_i, it lies in $Q_i^{+\epsilon}$: $\phi_{x_i}^{-1}Tz \in Q_i^{+\epsilon}$, or $Tz \in \phi_{x_i}Q_i^{+\epsilon}$. Hence (17.18) holds, which completes the proof. ■

Stieltjes Integrals

Suppose that F is a function on R^k satisfying the hypotheses of Theorem 12.5, so that there exists a measure μ such that $\mu(A) = \Delta_A F$ for bounded rectangles A. In integrals with respect to μ, $\mu(dx)$ is often replaced by $dF(x)$:

$$\int_A f(x)\,dF(x) = \int_A f(x)\mu(dx). \tag{17.22}$$

The left side of this equation is the *Stieltjes integral* of f with respect to F; since it is defined by the right side of the equation, nothing new is involved.

Suppose that f is uniformly continuous on a rectangle A, and suppose that A is decomposed into rectangles A_m small enough that $|f(x) - f(y)| < \epsilon/\mu(A)$ for $x, y \in A_m$. Then

$$\left| \int_A f(x)\,dF(x) - \sum_m f(x_m)\Delta_{A_m}F \right| < \epsilon$$

for $x_m \in A_m$. In this case the left side of (17.22) can be defined as the limit of these approximating sums without any reference to the general theory of measure, and for historical reasons it is sometimes called the *Riemann–Stieltjes* integral; (17.22) for the general f is then called the *Lebesgue–Stieltjes* integral. Since these distinctions are unimportant in the context of general measure theory, $\int f(x)\,dF(x)$ and $\int f\,dF$ are best regarded as merely notational variants for $\int f(x)\mu(dx)$ and $\int f\,d\mu$.

PROBLEMS

Let f be a bounded function on a bounded interval, say $[0, 1]$. Do not assume that f is a Borel function. Denote by L_*f and L^*f (L for Lebesgue) the lower and upper integrals as defined by (15.9) and (15.10), where μ is now Lebesgue measure λ on the Borel sets of $[0, 1]$. Denote by R_*f and R^*f (R for Riemann) the same quantities but with the outer supremum and infimum in (15.9) and (15.10) extending only over finite partitions of $[0, 1]$ into subintervals. It is obvious (see (15.11)) that

$$R_*f \leq L_*f \leq L^*f \leq R^*f. \tag{17.23}$$

Suppose that f is bounded, and consider these three conditions:

(i) *There is an r with the property that for each ϵ there is a δ such that* (17.1) *holds if $\{I_i\}$ partitions $[0,1]$ into subintervals with $\lambda(I_i) < \delta$ and if $x_i \in I_i$.*

(ii) $R_* f = R^* f$.

(iii) *If D_f is the set of points of discontinuity of f, then $\lambda(D_f) = 0$.*

The conditions are equivalent.

17.1. Prove:

(a) D_f is a Borel set.

(b) (i) implies (ii).

(c) (ii) implies (iii).

(d) (iii) implies (i).

(e) The r of (i) must coincide with the $R_* f = R^* f$ of (ii).

A note on terminology. An f on the general $(\Omega, \mathscr{F}, \mu)$ is defined to be integrable not if (15.12) holds, but if (16.1) does. And an f on $[0,1]$ is defined to be integrable with respect to Lebesgue measure not if $L_* f = L^* f$ holds, but, rather, if

$$\int_0^1 |f(x)|\,dx < \infty \tag{17.24}$$

does. The condition $L_* f = L^* f$ is not at issue, since for bounded f it always holds if f is a Borel function, and in this book f is always assumed to be a Borel function unless the contrary is explicitly stated. For the Lebesgue integral, the question is whether f is small enough that (17.24) holds, not whether it is sufficiently regular that $L_* f = L^* f$. For the Riemann integral, the terminology is different because $R_* f < R^* f$ holds for all sorts of important Borel functions, and one way to define Riemann integrability is to require $R_* f = R^* f$. In the context of general integration theory, one occasionally looks at the Riemann integral, but mostly for illustration and comparison.

17.2. 3.15 17.1↑ **(a)** Show that an indicator I_A for $A \subset [0,1]$ is Riemann integrable if and only if A is Jordan measurable.

(b) Find a Riemann integrable function that is not a Borel function.

17.3. Extend Theorem 17.1 to R^k.

17.4. Show that if f is integrable, then

$$\lim_{t \to 0} \int |f(x+t) - f(x)|\,dx = 0.$$

Use Theorem 17.1.

17.5. Suppose that μ is a finite measure on $\mathscr{R}^k$ and A is closed. Show that $\mu(x + A)$ is upper semicontinuous in x and hence measurable.

17.6. Suppose that $\int_0^\infty |f(x)|\,dx < \infty$. Show that for each ϵ, $\lambda[x\colon x > \alpha, |f(x)| > \epsilon] \to 0$ as $\alpha \to \infty$. Show by example that $f(x)$ need not go to 0 as $x \to \infty$ (even if f is continuous).

17.7. Let $f_n(x) = x^{n-1} - 2x^{2n-1}$. Calculate and compare $\int_0^1 \sum_{n=1}^{\infty} f_n(x)\,dx$ and $\sum_{n=1}^{\infty} \int_0^1 f_n(x)\,dx$. Relate this to Theorem 16.6 and to the corollary to Theorem 16.7.

17.8. Show that $(1+y^2)^{-1}$ has equal integrals over $(-\infty,-1),(-1,0),(0,1),(1,\infty)$. Conclude from (17.9) that $\int_0^1 (1+y^2)^{-1}\,dy = \pi/4$. Expand the integrand in a geometric series and deduce Leibniz's formula

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

by Theorem 16.7 (note that its corollary does not apply).

17.9. Show that if f is integrable, there exist continuous, integrable functions g_n such that $g_n(x) \to f(x)$ except on a set of Lebesgue measure 0. (Use Theorem 17.1(ii) with $\epsilon = n^{-2}$.)

17.10. 13.9 17.9↑ Let f be a finite-valued Borel function over $[0,1]$. By the following steps, prove *Lusin's theorem*: For each ϵ there exists a continuous function g such that $\lambda[x \in (0,1)\colon f(x) \neq g(x)] < \epsilon$.

(a) Show that f may be assumed integrable, or even bounded.

(b) Let g_n be continuous functions converging to f almost everywhere. Combine Egoroff's theorem and Theorem 12.3 to show that convergence is uniform on a compact set K such that $\lambda((0,1) - K) < \epsilon$. The limit $\lim_n g_n(x) = f(x)$ must be continuous when restricted to K.

(c) Exhibit $(0,1) - K$ as a disjoint union of open intervals I_k [A12]; define g as f on K, and define it by linear interpolation on each I_k.

17.11. Suppose in (17.7) that T' exists and is continuous and f is a Borel function, and suppose that $\int_a^b |f(Tx)T'(x)|\,dx < \infty$. Show in steps that $\int_{T(a,b)} |f(y)|\,dy < \infty$ and (17.8) holds. Prove this for (a) f continuous, (b) $f = I_{[s,t]}$, (c) $f = I_B$, (d) f simple, (e) $f \geq 0$, (f) f general.

17.12. 16.12↑ Let $\mathscr{L}$ consist of the continuous functions on R^1 with compact support. Show that $\mathscr{L}$ is a vector lattice in the sense of Problem 11.4 and has the property that $f \in \mathscr{L}$ implies $f \wedge 1 \in \mathscr{L}$ (note that $1 \notin \mathscr{L}$). Show that the σ-field $\mathscr{F}$ generated by $\mathscr{L}$ is $\mathscr{R}^1$. Suppose Λ is a positive linear functional on $\mathscr{L}$; show that Λ has the required continuity property if and only if $f_n(x) \downarrow 0$ uniformly in x implies $\Lambda(f_n) \to 0$. Show under this assumption on Λ that there is a measure μ on $\mathscr{R}^1$ such that

$$\Lambda(f) = \int f\,d\mu, \qquad f \in \mathscr{L}. \tag{17.25}$$

Show that μ is σ-finite and unique. This is a version of the *Riesz representation theorem*.

17.13. ↑ Let $\Lambda(f)$ be the Riemann integral of f, which does exist for f in $\mathscr{L}$. Using the most elementary facts about Riemann integration, show that the μ determined by (17.25) is Lebesgue measure. This gives still another way of constructing Lebesgue measure.

17.14. ↑ Extend the ideas in the preceding two problems to R^k.

SECTION 18. PRODUCT MEASURE AND FUBINI'S THEOREM

Let $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$ be measurable spaces. For given measures μ and ν on these spaces, the problem is to construct on the Cartesian product $X \times Y$ a *product measure* π such that $\pi(A \times B) = \mu(A)\nu(B)$ for $A \subset X$ and $B \subset Y$. In the case where μ and ν are Lebesgue measure on the line, π will be Lebesgue measure in the plane. The main result is *Fubini's theorem*, according to which double integrals can be calculated as iterated integrals.

Product Spaces

It is notationally convenient in this section to change from $(\Omega, \mathscr{F})$ to $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$. In the product space $X \times Y$ a *measurable rectangle* is a product $A \times B$ for which $A \in \mathscr{X}$ and $B \in \mathscr{Y}$. The natural class of sets in $X \times Y$ to consider is the σ-field $\mathscr{X} \times \mathscr{Y}$ generated by the measurable rectangles. (Of course, $\mathscr{X} \times \mathscr{Y}$ is not a Cartesian product in the usual sense.)

Example 18.1. Suppose that $X = Y = R^1$ and $\mathscr{X} = \mathscr{Y} = \mathscr{R}^1$. Then a measurable rectangle is a Cartesian product $A \times B$ in which A and B are linear Borel sets. The term *rectangle* has up to this point been reserved for Cartesian products of intervals, and so a measurable rectangle is more general. As the measurable rectangles do include the ordinary ones and the latter generate $\mathscr{R}^2$, it follows that $\mathscr{R}^2 \subset \mathscr{R}^1 \times \mathscr{R}^1$. On the other hand, if A is an interval, $[B\colon A \times B \in \mathscr{R}^2]$ contains R^1 $(A \times R^1 = \bigcup_n (A \times (-n, n]) \in \mathscr{R}^2)$ and is closed under the formation of proper differences and countable unions; thus it is a σ-field containing the intervals and hence the Borel sets. Therefore, if B is a Borel set, $[A\colon A \times B \in \mathscr{R}^2]$ contains the intervals and hence, being a σ-field, contains the Borel sets. Thus all the measurable rectangles are in $\mathscr{R}^2$, and so $\mathscr{R}^1 \times \mathscr{R}^1 = \mathscr{R}^2$ consists exactly of the two-dimensional Borel sets. ■

As this example shows, $\mathscr{X} \times \mathscr{Y}$ is in general much larger than the class of measurable rectangles.

Theorem 18.1. (i) *If $E \in \mathscr{X} \times \mathscr{Y}$, then for each x the set $[y\colon (x, y) \in E]$ lies in $\mathscr{Y}$ and for each y the set $[x\colon (x, y) \in E]$ lies in $\mathscr{X}$.*

(ii) *If f is measurable $\mathscr{X} \times \mathscr{Y}$, then for each fixed x the function $f(x, \cdot)$ is measurable $\mathscr{Y}$, and for each fixed y the function $f(\cdot, y)$ is measurable $\mathscr{X}$.*

The set $[y\colon (x, y) \in E]$ is the *section* of E determined by x, and $f(x, \cdot)$ is the *section* of f determined by x.

PROOF. Fix x, and consider the mapping $T_x\colon Y \to X \times Y$ defined by $T_x y = (x, y)$. If $E = A \times B$ is a measurable rectangle, $T_x^{-1}E$ is B or $\varnothing$

according as A contains x or not, and in either case $T_x^{-1}E \in \mathscr{Y}$. By Theorem 13.1(i), T_x is measurable $\mathscr{Y}/\mathscr{X}\times\mathscr{Y}$. Hence $[y\colon (x,y)\in E] = T_x^{-1}E \in \mathscr{Y}$ for $E \in \mathscr{X}\times\mathscr{Y}$. By Theorem 13.1(ii), if f is measurable $\mathscr{X}\times\mathscr{Y}/\mathscr{R}^1$, then fT_x is measurable $\mathscr{Y}/\mathscr{R}^1$. Hence $f(x,\cdot) = fT_x(\cdot)$ is measurable $\mathscr{Y}$. The symmetric statements for fixed y are proved the same way. ■

Product Measure

Now suppose that $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$ are measure spaces, and suppose for the moment that μ and ν are *finite*. By the theorem just proved $\nu[y\colon (x,y)\in E]$ is a well-defined function of x. If $\mathscr{L}$ is the class of E in $\mathscr{X}\times\mathscr{Y}$ for which this function is measurable $\mathscr{X}$, it is not hard to show that $\mathscr{L}$ is a λ-system. Since the function is $I_A(x)\nu(B)$ for $E = A\times B$, $\mathscr{L}$ contains the π-system consisting of the measurable rectangles. Hence $\mathscr{L}$ coincides with $\mathscr{X}\times\mathscr{Y}$ by the π-λ theorem. It follows without difficulty that

$$\pi'(E) = \int_X \nu[y\colon (x,y)\in E]\,\mu(dx), \qquad E \in \mathscr{X}\times\mathscr{Y}, \tag{18.1}$$

is a finite measure on $\mathscr{X}\times\mathscr{Y}$, and similarly for

$$\pi''(E) = \int_Y \mu[x\colon (x,y)\in E]\,\nu(dy), \qquad E \in \mathscr{X}\times\mathscr{Y}. \tag{18.2}$$

For measurable rectangles,

$$\pi'(A\times B) = \pi''(A\times B) = \mu(A)\cdot\nu(B). \tag{18.3}$$

The class of E in $\mathscr{X}\times\mathscr{Y}$ for which $\pi'(E) = \pi''(E)$ thus contains the measurable rectangles; since this class is a λ-system, it contains $\mathscr{X}\times\mathscr{Y}$. The common value $\pi'(E) = \pi''(E)$ is the product measure sought.

To show that (18.1) and (18.2) also agree for σ-finite μ and ν, let $\{A_m\}$ and $\{B_n\}$ be decompositions of X and Y into sets of finite measure, and put $\mu_m(A) = \mu(A\cap A_m)$ and $\nu_n(B) = \nu(B\cap B_n)$. Since $\nu(B) = \sum_m \nu_m(B)$, the integrand in (18.1) is measurable $\mathscr{X}$ in the σ-finite as well as in the finite case; hence π' is a well-defined measure on $\mathscr{X}\times\mathscr{Y}$ and so is π''. If π'_{mn} and π''_{mn} are (18.1) and (18.2) for μ_m and ν_n, then by the finite case, already treated (see Example 16.5), $\pi'(E) = \sum_{mn}\pi'_{mn}(E) = \sum_{mn}\pi''_{mn}(E) = \pi''(E)$. Thus (18.1) and (18.2) coincide in the σ-finite case as well. Moreover, $\pi'(A\times B) = \sum_{mn}\mu_m(A)\nu_n(B) = \mu(A)\nu(B)$.

Theorem 18.2. *If $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$ are σ-finite measure spaces, $\pi(E) = \pi'(E) = \pi''(E)$ defines a σ-finite measure on $\mathscr{X}\times\mathscr{Y}$; it is the only measure such that $\pi(A\times B) = \mu(A)\cdot\nu(B)$ for measurable rectangles.*

PROOF. Only σ-finiteness and uniqueness remain to be proved. The products $A_m \times B_n$ for $\{A_m\}$ and $\{B_n\}$ as above decompose $X \times Y$ into measurable rectangles of finite π-measure. This proves both σ-finiteness and uniqueness, since the measurable rectangles form a π-system generating $\mathscr{X} \times \mathscr{Y}$ (Theorem 10.3). ■

The π thus defined is called *product measure*; it is usually denoted $\mu \times \nu$. Note that the integrands in (18.1) and (18.2) may be infinite for certain x and y, which is one reason for introducing functions with infinite values. Note also that (18.3) in some cases requires the conventions (15.2).

Fubini's Theorem

Integrals with respect to π are usually computed via the formulas

$$\int_{X\times Y} f(x,y)\pi(d(x,y)) = \int_X \left[\int_Y f(x,y)\nu(dy)\right]\mu(dx) \tag{18.4}$$

and

$$\int_{X\times Y} f(x,y)\pi(d(x,y)) = \int_Y \left[\int_X f(x,y)\mu(dx)\right]\nu(dy). \tag{18.5}$$

The left side here is a *double* integral, and the right sides are *iterated* integrals. The formulas hold very generally, as the following argument shows.

Consider (18.4). The inner integral on the right is

$$\int_Y f(x,y)\nu(dy). \tag{18.6}$$

Because of Theorem 18.1(ii), for f measurable $\mathscr{X} \times \mathscr{Y}$ the integrand here is measurable $\mathscr{Y}$; the question is whether the integral exists, whether (18.6) is measurable $\mathscr{X}$ as a function of x, and whether it integrates to the left side of (18.4).

First consider nonnegative f. If $f = I_E$, everything follows from Theorem 18.2: (18.6) is $\nu[y\colon (x,y) \in E]$, and (18.4) reduces to $\pi(E) = \pi'(E)$. Because of linearity (Theorem 15.1(iv)), if f is a nonnegative simple function, then (18.6) is a linear combination of functions measurable $\mathscr{X}$ and hence is itself measurable $\mathscr{X}$; further application of linearity to the two sides of (18.4) shows that (18.4) again holds. The general nonnegative f is the monotone limit of nonnegative simple functions; applying the monotone convergence theorem to (18.6) and then to each side of (18.4) shows that again f has the properties required.

Thus for nonnegative f, (18.6) is a well-defined function of x (the value ∞ is not excluded), measurable $\mathscr{X}$, whose integral satisfies (18.4). If one side of

(18.4) is infinite, so is the other; if both are finite, they have the same finite value.

Now suppose that f, not necessarily nonnegative, is integrable with respect to π. Then the two sides of (18.4) are finite if f is replaced by $|f|$. Now make the further assumption that

$$\int_Y |f(x,y)|\nu(dy) < \infty \tag{18.7}$$

for all x. Then

$$\int_Y f(x,y)\nu(dy) = \int_Y f^+(x,y)\nu(dy) - \int_Y f^-(x,y)\nu(dy). \tag{18.8}$$

The functions on the right here are measurable $\mathscr{X}$ and (since $f^+, f^- \le |f|$) integrable with respect to μ, and so the same is true of the function on the left. Integrating out the x and applying (18.4) to f^+ and to f^- gives (18.4) for f itself.

The set A_0 of x satisfying (18.7) need not coincide with X, but $\mu(X - A_0) = 0$ if f is integrable with respect to π, because the function in (18.7) integrates to $\int |f|\, d\pi$ (Theorem 15.2(iii)). Now (18.8) holds on A_0, (18.6) is measurable $\mathscr{X}$ on A_0, and (18.4) again follows if the inner integral on the right is given some arbitrary constant value on $X - A_0$.

The same analysis applies to (18.5):

Theorem 18.3. *Under the hypotheses of Theorem* 18.2, *for nonnegative* f *the functions*

$$\int_Y f(x,y)\nu(dy), \int_X f(x,y)\mu(dx) \tag{18.9}$$

are measurable $\mathscr{X}$ *and* $\mathscr{Y}$, *respectively, and* (18.4) *and* (18.5) *hold. If* f (*not necessarily nonnegative*) *is integrable with respect to* π, *then the two functions* (18.9) *are finite and measurable on* A_0 *and on* B_0, *respectively, where* $\mu(X - A_0) = \nu(Y - B_0) = 0$, *and again* (18.4) *and* (18.5) *hold.*

It is understood here that the inner integrals on the right in (18.4) and (18.5) are set equal to 0 (say) outside A_0 and B_0.[†]

This is *Fubini's theorem*; the part concerning nonnegative f is sometimes called *Tonelli's theorem*. Application of the theorem usually follows a two-step procedure that parallels its proof. First, one of the iterated integrals is computed (or estimated above) with $|f|$ in place of f. If the result is finite,

[†]Since two functions that are equal almost everywhere have the same integral, the theory of integration could be extended to functions that are only *defined* almost everywhere; then A_0 and B_0 would disappear from Theorem 18.3.

then the double integral (integral with respect to π) of $|f|$ must be finite, so that f is integrable with respect to π; then the value of the double integral of f is found by computing one of the iterated integrals of f. If the iterated integral of $|f|$ is infinite, f is not integrable π.

Example 18.2. Let D_r be the closed disk in the plane with center at the origin and radius r. By (17.12),

$$\lambda_2(D_r) = \iint_{D_r} dx\,dy = \iint_{\substack{0<\rho\le r \\ 0<0<2\pi}} \rho\,d\rho\,d\theta.$$

The last integral can be evaluated by Fubini's theorem:

$$\lambda_2(D_r) = 2\pi\int_0^r \rho\,d\rho = \pi r^2. \qquad \blacksquare$$

Example 18.3. Let $I = \int_{-\infty}^{\infty} e^{-x^2}\,dx$. By Fubini's theorem applied in the plane and by the polar-coordinate formula,

$$I^2 = \iint_{R^2} e^{-(x^2+y^2)}\,dx\,dy = \iint_{\substack{\rho>0 \\ 0<\theta<2\pi}} e^{-\rho^2}\rho\,d\rho\,d\theta.$$

The double integral on the right can be evaluated as an iterated integral by another application of Fubini's theorem, which leads to the famous formula

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}\,. \tag{18.10}$$

As the integrand in this example is nonnegative, the question of integrability does not arise. $\blacksquare$

Example 18.4. It is possible by means of Fubini's theorem to identify the limit in (17.4). First,

$$\int_0^t e^{-ux}\sin x\,dx = \frac{1}{1+u^2}\left[1 - e^{-ut}(u\sin t + \cos t)\right],$$

as follows by differentiation with respect to t. Since

$$\int_0^t\left[\int_0^\infty |e^{-ux}\sin x|\,du\right]dx = \int_0^t |\sin x|\cdot x^{-1}\,dx \le t < \infty,$$

Fubini's theorem applies to the integration of $e^{-ux}\sin x$ over $(0,t)\times(0,\infty)$:

$$\int_0^t \frac{\sin x}{x}dx = \int_0^t \sin x\left[\int_0^\infty e^{-ux}\,du\right]dx$$

$$= \int_0^\infty\left[\int_0^t e^{-ux}\sin x\,dx\right]du$$

$$= \int_0^\infty \frac{du}{1+u^2} - \int_0^\infty \frac{e^{-ut}}{1+u^2}(u\sin t+\cos t)\,du.$$

The next-to-last integral is $\pi/2$ (see (17.9)), and a change of variable $ut = s$ shows that the final integral goes to 0 as $t\to\infty$. Therefore,

(18.11) $$\lim_{t\to\infty}\int_0^t \frac{\sin x}{x}dx = \frac{\pi}{2}.$$ ■

Integration by Parts

Let F and G be two nondecreasing, right-continuous functions on an interval $[a,b]$, and let μ and ν be the corresponding measures:

$$\mu(x,y] = F(y)-F(x),\quad \nu(x,y] = G(y)-G(x),\qquad a\le x\le y\le b.$$

In accordance with the convention (17.22) write $dF(x)$ and $dG(x)$ in place of $\mu(dx)$ and $\nu(dx)$.

Theorem 18.4. *If F and G have no common points of discontinuity in $(a,b]$, then*

(18.12) $$\int_{(a,b]} G(x)\,dF(x)$$

$$= F(b)G(b) - F(a)G(a) - \int_{(a,b]} F(x)\,dG(x).$$

In brief: $\int G\,dF = \Delta FG - \int F\,dG$. This is one version of the partial integration formula.

PROOF. Note first that replacing $F(x)$ by $F(x)-C$ leaves (18.12) unchanged—it merely adds and subtracts $C\nu(a,b]$ on the right. Hence (take $C = F(a)$) it is no restriction to assume that $F(x) = \mu(a,x]$ and no restriction to assume that $G(x) = \nu(a,x]$. If $\pi = \mu\times\nu$ is product measure in the plane,

then by Fubini's theorem,

$$(18.13)\qquad \pi[(x,y): a<y\le x\le b] = \int_{(a,b]} \nu(a,x]\mu(dx) = \int_{(a,b]} G(x)\,dF(x)$$

and

$$(18.14)\qquad \pi[(x,y): a<x\le y\le b] = \int_{(a,b]} \mu(a,y]\nu(dy) = \int_{(a,b]} F(y)\,dG(y).$$

The two sets on the left have as their union the square $S=(a,b]\times(a,b]$. The diagonal of S has π-measure

$$\pi[(x,y): a<x=y\le b] = \int_{(a,b]} \nu\{x\}\mu(dx) = 0$$

because of the assumption that μ and ν share no points of positive measure. Thus the left sides of (18.13) and (18.14) add to $\pi(S)=\mu(a,b]\nu(a,b] = F(b)G(b)$. ■

Suppose that ν has a density g with respect to Lebesgue measure and let $G(x)=c+\int_a^x g(t)\,dt$. Transform the right side of (18.12) by the formula (16.13) for integration with respect to a density; the result is

$$(18.15)\qquad \int_{(a,b]} G(x)\,dF(x) = F(b)G(b) - F(a)G(a) - \int_a^b F(x)g(x)\,dx.$$

A consideration of positive and negative parts shows that this holds for any g integrable over $(a,b]$.

Suppose further that μ has a density f with respect to Lebesgue measure, and let $F(x)=c'+\int_a^x f(t)\,dt$. Then (18.15) further reduces to

$$(18.16)\qquad \int_a^b G(x)f(x)\,dx = F(b)G(b) - F(a)G(a) - \int_a^b F(x)g(x)\,dx.$$

Again, f can be any integrable function. This is the classical formula for integration by parts.

Under the appropriate integrability conditions, $(a,b]$ can be replaced by an unbounded interval.

Products of Higher Order

Suppose that $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \nu)$, and $(Z, \mathscr{Z}, \eta)$ are three σ-finite measure spaces. In the usual way, identify the products $X \times Y \times Z$ and $(X \times Y) \times Z$. Let $\mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$ be the σ-field in $X \times Y \times Z$ generated by the $A \times B \times C$ with A, B, C in $\mathscr{X}, \mathscr{Y}, \mathscr{Z}$, respectively. For C in Z, let $\mathscr{G}_C$ be the class of $E \in \mathscr{X} \times \mathscr{Y}$ for which $E \times C \in \mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$. Then $\mathscr{G}_C$ is a σ-field containing the measurable rectangles in $X \times Y$, and so $\mathscr{G}_C = \mathscr{X} \times \mathscr{Y}$. Therefore, $(\mathscr{X} \times \mathscr{Y}) \times \mathscr{Z} \subset \mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$. But the reverse relation is obvious, and so $(\mathscr{X} \times \mathscr{Y}) \times \mathscr{Z} = \mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$.

Define the product $\mu \times \nu \times \eta$ on $\mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$ as $(\mu \times \nu) \times \eta$. It gives to $A \times B \times C$ the value $(\mu \times \nu)(A \times B) \cdot \eta(C) = \mu(A)\nu(B)\eta(C)$, and it is the only measure that does. The formulas (18.4) and (18.5) extend in the obvious way.

Products of four or more components can clearly be treated in the same way. This leads in particular to another construction of Lebesgue measure in $R^k = R^1 \times \cdots \times R^1$ (see Example 18.1) as the product $\lambda \times \cdots \times \lambda$ (k factors) on $\mathscr{R}^k = \mathscr{R}^1 \times \cdots \times \mathscr{R}^1$. Fubini's theorem of course gives a practical way to calculate volumes:

Example 18.5. Let V_k be the volume of the sphere of radius 1 in R^k; by Theorem 12.2, a sphere in R^k with radius r has volume $r^k V_k$. Let A be the unit sphere in R^k, let $B = [(x_1, x_2)\colon\ x_1^2 + x_2^2 \le 1]$, and let $C(x_1, x_2) = [(x_3, \ldots, x_k)\colon \sum_{i=3}^k \ldots x_i^2 \le 1 - x_1^2 - x_2^2]$. By Fubini's theorem,

$$V_k = \int_A dx_1 \cdots dx_k = \int_B dx_1\, dx_2 \int_{C(x_1, x_2)} dx_3 \cdots dx_k$$

$$= \int_B dx_1\, dx_2\, V_{k-2}\left(1 - x_1^2 - x_2^2\right)^{(k-2)/2}$$

$$= V_{k-2} \iint\limits_{\substack{0<\theta<2\pi \\ 0<\rho<1}} \left(1 - \rho^2\right)^{(k-2)/2} \rho\, d\rho\, d\theta$$

$$= \pi V_{k-2} \int_0^1 t^{(k-2)/2}\, dt = \frac{2\pi V_{k-2}}{k}.$$

If V_0 is taken as 1, this holds for $k = 2$ as well as for $k \ge 3$. Since $V_1 = 2$, it follows by induction that

$$V_{2i-1} = \frac{2(2\pi)^{i-1}}{1 \times 3 \times 5 \times \cdots \times (2i-1)}, \qquad V_{2i} = \frac{(2\pi)^i}{2 \times 4 \times \cdots \times (2i)}$$

for $i = 1, 2, \ldots$. Example 18.2 is a special case. ∎

PROBLEMS

18.1. Show by Theorem 18.1 that if $A \times B$ is nonempty and lies in $\mathscr{X} \times \mathscr{Y}$, then $A \in \mathscr{X}$ and $B \in \mathscr{Y}$.

18.2. 2.9↑ Suppose that $X = Y$ is uncountable and $\mathscr{X} = \mathscr{Y}$ consists of the countable and the cocountable sets. Show that the diagonal $E = [(x, y): x = y]$ does not lie in $\mathscr{X} \times \mathscr{Y}$, even though $[y: (x, y) \in E] \in \mathscr{Y}$ and $[x: (x, y) \in E] \in \mathscr{X}$ for all x and y.

18.3. 10.5 18.1↑ Let $(X, \mathscr{X}, \mu) = (Y, \mathscr{Y}, \nu)$ be the completion of $(R^1, \mathscr{R}^1, \lambda)$. Show that $(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu \times \nu)$ is not complete.

18.4. The assumption of σ-finiteness in Theorem 18.2 is essential: Let μ be Lebesgue measure on the line, let ν be counting measure on the line, and take $E = [(x, y): x = y]$. Then (18.1) and (18.2) do not agree.

18.5. Example 18.2 in effect connects π as the area of the unit disk D_1 with the π of trigonometry.

(a) A second way: Calculate $\lambda_2(D_1)$ directly by Fubini's theorem: $\lambda_2(D_1) = \int_{-1}^{1} 2(1 - x^2)^{1/2}\, dx$. Evaluate the integral by trigonometric substitution.

(b) A third way: Inscribe in the unit circle a regular polygon of n sides. Its interior consists of n congruent isosceles triangles with angle $2\pi/n$ at the apex; the area is $n \sin(\pi/n)\cos(\pi/n)$, which goes to π.

18.6. Suppose that f is nonnegative on a σ-finite measure space $(\Omega, \mathscr{F}, \mu)$. Show that

$$\int_{\Omega} f\, d\mu = (\mu \times \lambda)\big[(\omega, y) \in \Omega \times R^1 : 0 \le y \le f(\omega)\big].$$

Prove that the set on the right is measurable. This gives the "area under the curve." Given the existence of $\mu \times \lambda$ on $\Omega \times R^1$, one can use the right side of this equation as an alternative definition of the integral.

18.7. Reconsider Problem 12.12.

18.8. Suppose that $\nu[y: (x, y) \in E] = \nu[y: (x, y) \in F]$ for all x, and show that $(\mu \times \nu)(E) = (\mu \times \nu)(F)$. This is a general version of *Cavalieri's principle*.

18.9. **(a)** Suppose that μ is σ-finite, and prove the corollary to Theorem 16.7 by Fubini's theorem in the product of $(\Omega, \mathscr{F}, \mu)$ and $\{1, 2, \ldots\}$ with counting measure.

(b) Relate the series in Problem 17.7 to Fubini's theorem.

18.10. (a) Let $\mu = \nu$ be counting measure on $X = Y = \{1, 2, \ldots\}$. If

$$f(x,y) = \begin{cases} 2 - 2^{-x} & \text{if } x = y, \\ -2 + 2^{-x} & \text{if } x = y + 1, \\ 0 & \text{otherwise,} \end{cases}$$

then the iterated integrals exist but are unequal. Why does this not contradict Fubini's theorem?

(b) Show that $xy/(x^2 + y^2)^2$ is not integrable over the square $[(x, y): |x|, |y| \le 1]$ even though the iterated integrals exist and are equal.

18.11. Exhibit a case in which (18.12) fails because F and G have a common point of discontinuity.

18.12. Prove (18.16) for the case in which all the functions are continuous by differentiating with respect to the upper limit of integration.

18.13. Prove for distribution functions F that $\int_{-\infty}^{\infty}(F(x + c) - F(x))\,dx = c$.

18.14. Prove for continuous distribution functions that $\int_{-\infty}^{\infty} F(x)\,dF(x) = \frac{1}{2}$.

18.15. Suppose that a number f_n is defined for each $n \ge n_0$ and put $F(x) = \sum_{n_0 \le n \le x} f_n$. Deduce from (18.15) that

$$\text{(18.17)} \qquad \sum_{n_0 \le n \le x} G(n) f_n = F(x)G(x) - \int_{n_0}^{x} F(t)g(t)\,dt$$

if $G(y) - G(x) = \int_x^y g(t)\,dt$, which will hold if G has continuous derivative g. First assume that the f_n are nonnegative.

18.16. ↑ Take $n_0 = 1$, $f_n = 1$, and $G(x) = 1/x$, and derive $\sum_{n \le x} n^{-1} = \log x + \gamma + O(1/x)$, where $\gamma = 1 - \int_1^{\infty}(t - \lfloor t \rfloor)t^{-2}\,dt$ is Euler's constant.

18.17. 5.20 18.15↑ Use (18.17) and (5.51) to prove that there exists a constant c such that

$$\text{(18.18)} \qquad \sum_{p \le x} \frac{1}{p} = \log\log x + c + O\left(\frac{1}{\log x}\right).$$

18.18. Euler's *gamma function* is defined for positive t by $\Gamma(t) = \int_0^{\infty} x^{t-1}e^{-x}\,dx$.

(a) Prove that $\Gamma^{(k)}(t) = \int_0^{\infty} x^{t-1}(\log x)^k e^{-x}\,dx$.

(b) Show by partial integration that $\Gamma(t + 1) = t\Gamma(t)$ and hence that $\Gamma(n + 1) = n!$ for integral n.

(c) From (18.10) deduce $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

(d) Show that the unit sphere in R^k has volume (see Example 18.5)

$$\text{(18.19)} \qquad V_k = \frac{\pi^{k/2}}{\Gamma((k/2) + 1)}.$$

18.19. By partial integration prove that $\int_0^\infty (\sin x)/x)^2\,dx = \pi/2$ and $\int_{-\infty}^{\infty}(1-\cos x)x^{-2}\,dx = \pi$.

18.20. Suppose that μ is a probability measure on $(X, \mathscr{X})$ and that, for each x in X, ν_x is a probability measure on $(Y, \mathscr{Y})$. Suppose further that, for each B in $\mathscr{Y}$, $\nu_x(B)$ is, as a function of x, measurable $\mathscr{X}$. Regard the $\mu(A)$ as initial probabilities and the $\nu_x(B)$ as transition probabilities.

(a) Show that, if $E \in \mathscr{X} \times \mathscr{Y}$, then $\nu_x[y\colon (x, y) \in E]$ is measurable $\mathscr{X}$.

(b) Show that $\pi(E) = \int_X \nu_x[y\colon (x, y) \in E]\mu(dx)$ defines a probability measure on $\mathscr{X} \times \mathscr{Y}$. If $\nu_x = \nu$ does not depend on x, this is just (18.1).

(c) Show that if f is measurable $\mathscr{X} \times \mathscr{Y}$ and nonnegative, then $\int_Y f(x, y)\nu_x(dy)$ is measurable $\mathscr{X}$. Show further that

$$\int_{X\times Y} f(x, y)\pi(d(x, y)) = \int_X \left[\int_Y f(x, y)\nu_x(dy)\right]\mu(dx),$$

which extends Fubini's theorem (in the probability case). Consider also f's that may be negative.

(d) Let $\nu(B) = \int_X \nu_x(B)\mu(dx)$. Show that $\pi(X \times B) = \nu(B)$ and

$$\int_Y f(y)\nu(dy) = \int_X \left[\int_Y f(y)\nu_x(dy)\right]\mu(dx).$$

SECTION 19. THE L^p SPACES*

Definitions

Fix a measure space $(\Omega, \mathscr{F}, \mu)$. For $1 \le p < \infty$, let $L^p = L^p(\Omega, \mathscr{F}, \mu)$ be the class of (measurable) real functions f for which $|f|^p$ is integrable, and for f in L^p, write

$$\|f\|_p = \left[\int |f|^p\,d\mu\right]^{1/p}. \tag{19.1}$$

There is a special definition for the case $p = \infty$: The *essential supremum* of f is

$$\|f\|_\infty = \inf[\alpha\colon \mu[\omega\colon |f(\omega)| > \alpha] = 0]; \tag{19.2}$$

take L^∞ to consist of those f for which this is finite. The spaces L^p have a geometric structure that can usefully guide the intuition. The basic facts are laid out in this section, together with two applications to theoretical statistics.

*The results in this section are used nowhere else in the book. The proofs require some elementary facts about metric spaces, vector spaces, and convex sets; and in one place the Radon–Nikodym theorem of Section 32 is used. As a matter of fact, it is possible to jump ahead and read Section 32 at this point, since it makes no use of Chapters 4 and 5.

For $1 < p, q < \infty$, p and q are *conjugate* indices if $p^{-1} + q^{-1} = 1$; p and q are also conjugate if one is 1 and the other is ∞ (formally, $1^{-1} + \infty^{-1} = 1$). *Hölder's inequality* says that if p and q are conjugate and if $f \in L^p$ and $g \in L^q$, then fg is integrable and

$$\left| \int fg \, d\mu \right| \leq \int |fg| \, d\mu \leq \|f\|_p \|g\|_q. \tag{19.3}$$

This is obvious if $p = 1$ and $q = \infty$. If $1 < p$, $q < \infty$ and μ is a probability measure, and if f and g are simple functions, (19.3) is (5.35). But the proof in Section 5 goes over without change to the general case.

Minkowski's inequality says that if $f, g \in L^p$ $(1 \leq p \leq \infty)$, then $f + g \in L^p$ and

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{19.4}$$

This is clear if $p = 1$ or $p = \infty$. Suppose that $1 < p < \infty$. Since $|f + g| \leq 2(|f|^p + |g|^p)^{1/p}$, $f + g$ does lie in L^p. Let q be conjugate to p. Since $p - 1 = p/q$, Hölder's inequality gives

$$\begin{aligned}
\|f + g\|_p^p &= \int |f + g|^p \, d\mu \\
&\leq \int |f| \cdot |f + g|^{p/q} \, d\mu + \int |g| \cdot |f + g|^{p/q} \, d\mu \\
&\leq \|f\|_p \cdot \| |f + g|^{p/q} \|_q + \|g\|_p \cdot \| |f + g|^{p/q} \|_q \\
&= (\|f\|_p + \|g\|_p) \left[\int |f + g|^p \, d\mu \right]^{1/q} \\
&= (\|f\|_p + \|g\|_p) \|f + g\|_p^{p/q}.
\end{aligned}$$

Since $p - p/q = 1$, (19.4) follows.†

If α is real and $f \in L^p$, then obviously $\alpha f \in L^p$ and

$$\|\alpha f\|_p = |\alpha| \cdot \|f\|_p. \tag{19.5}$$

Define a metric on L^p by $d_p(f, g) = \|f - g\|_p$. Minkowski's inequality gives the triangle inequality for d_p, and d_p is certainly symmetric. Further, $d_p(f, g) = 0$ if and only if $|f - g|^p$ integrates to 0, that is, $f = g$ almost everywhere. To make L^p a metric space, identify functions that are equal almost everywhere.

†The Hölder and Minkowski inequalities can also be proved by convexity arguments; see Problems 5.9 and 5.10.

If $\|f-f_n\|_p \to 0$ and $p<\infty$, so that $\int|f-f_n|^p\,d\mu \to 0$, then f_n is said to converge to f *in the mean of order* p.

If $f=f'$ and $g=g'$ almost everywhere, then $f+g=f'+g'$ almost everywhere, and for real α, $\alpha f=\alpha f'$ almost everywhere. In L^p, f and f' become the same function, and similarly for the pairs g and g', $f+g$ and $f'+g'$, and αf and $\alpha f'$. This means that addition and scalar multiplication are unambiguously defined in L^p, which is thus a real vector space. It is a *normed vector space* in the sense that it is equipped with a *norm* $\|\cdot\|_p$ satisfying (19.4) and (19.5).

Completeness and Separability

A normed vector space is a *Banach* space if it is complete under the corresponding metric. According to the *Riesz–Fischer theorem*, this is true of L^p:

Theorem 19.1. *The space L^p is complete.*

PROOF. It is enough to show that every fundamental sequence contains a convergent subsequence. Suppose first that $p<\infty$. Assume that $\|f_m-f_n\|_p \to 0$ as $m,n\to\infty$, and choose an increasing sequence $\{n_k\}$ so that $\|f_m-f_n\|_p^p \le 2^{-(p+1)k}$ for $m,n\ge n_k$. Since $\int|f_m-f_n|^p\,d\mu \ge \alpha^p\mu[|f_m-f_n|\ge\alpha]$ (this is just a general version of Markov's inequality (5.31)), $\mu[|f_n-f_m|\ge 2^{-k}]\le 2^{pk}\|f_m-f_n\|_p^p \le 2^{-k}$ for $m,n\ge n_k$. Therefore, $\sum_k\mu[|f_{n_{k+1}}-f_{n_k}|\ge 2^{-k}]<\infty$, and it follows by the first Borel–Cantelli lemma (which works for arbitrary measures) that, outside a set of μ-measure 0, $\sum_k|f_{n_{k+1}}-f_{n_k}|$ converges. But then f_{n_k} converges to some f almost everywhere, and by Fatou's lemma, $\int|f-f_{n_k}|^p\,d\mu \le \liminf_i \int|f_{n_i}-f_{n_k}|^p\,d\mu \le 2^{-k}$. Therefore, $f\in L^p$ and $\|f-f_{n_k}\|_p\to 0$, as required.

If $p=\infty$, choose $\{n_k\}$ so that $\|f_m-f_n\|_\infty \le 2^{-k}$ for $m,n\ge n_k$. Since $|f_{n_{k+1}}-f_{n_k}|\le 2^{-k}$ almost everywhere, f_{n_k} converges to some f, and $|f-f_{n_k}|\le 2^{-k}$ almost everywhere. Again, $\|f-f_{n_k}\|_p\to 0$. ∎

The next theorem has to do with separability.

Theorem 19.2. (i) *Let U be the set of simple functions $\sum_{i=1}^m \alpha_i I_{B_i}$ for α_i and $\mu(B_i)$ finite. For $1\le p\le\infty$, U is dense in L^p.*

(ii) *If μ is σ-finite and $\mathscr{F}$ is countably generated, and if $p<\infty$, then L^p is separable.*

PROOF. *Proof of* (i). Suppose first that $p<\infty$. For $f\in L^p$, choose (Theorem 13.5) simple functions f_n such that $f_n\to f$ and $|f_n|\le|f|$. Then $f_n\in L^p$, and by the dominated convergence theorem, $\int|f-f_n|^p\,d\mu\to 0$. Therefore,

$\|f-f_n\|_p<\epsilon$ for some n; but each f_n is in U. As for the case $p=\infty$, if $n2^n>\|f\|_\infty$, then the f_n defined by (13.6) satisfies $\|f-f_n\|_\infty\le 2^{-n}$ ($<\epsilon$ for large n).

Proof of (*ii*). Suppose that $\mathscr{F}$ is generated by a countable class $\mathscr{C}$ and that Ω is covered by a countable class $\mathscr{D}$ of $\mathscr{F}$-sets of finite measure. Let $E_1, E_2, \ldots$ be an enumeration of $\mathscr{C}\cup\mathscr{D}$; let $\mathscr{P}_n$ ($n\ge 1$) be the partition consisting of the sets of the form $F_1\cap\cdots\cap F_n$, where F_i is E_i or E_i^c; and let $\mathscr{F}_n$ be the field of unions of the sets in $\mathscr{P}_n$. Then $\mathscr{F}_0=\bigcup_{n=1}^\infty\mathscr{F}_n$ is a countable field that generates $\mathscr{F}$, and μ is σ-finite on $\mathscr{F}_0$. Let V be the set of simple functions $\sum_{i=1}^m\alpha_i I_{A_i}$ for α_i rational, $A_i\in\mathscr{F}_0$, and $\mu(A_i)<\infty$.

Let $g=\sum_{i=1}^m\alpha_i I_{B_i}$ be the element of U constructed in the proof of part (i). Then $\|f-g\|_p<\epsilon$, the α_i are rational by (13.6), and any α_i that vanish can be suppressed. By Theorem 11.4(ii), there exist sets A_i in $\mathscr{F}_0$ such that $\mu(B_i\Delta A_i)<(\epsilon/m|\alpha_i|)^p$, provided $p<\infty$, and then $h=\sum_{i=1}^m\alpha_i I_{A_i}$ lies in V and $\|f-h\|_p<2\epsilon$. But V is countable.† ■

Conjugate Spaces

A *linear functional* on L^p is a real-valued function γ such that

$$\gamma(\alpha f+\alpha' f')=\alpha\gamma(f)+\alpha'\gamma(f'). \tag{19.6}$$

The functional is *bounded* if there is a finite M such that

$$|\gamma(f)|\le M\cdot\|f\|_p \tag{19.7}$$

for all f in L^p. A bounded linear functional is uniformly continuous on L^p because $\|f-f'\|_p<\epsilon/M$ implies $|\gamma(f)-\gamma(f')|<\epsilon$ (if $M>0$; and $M=0$ implies $\gamma(f)\equiv 0$). The *norm* $\|\gamma\|$ of γ is the smallest M that works in (19.7): $\|\gamma\|=\sup[|\gamma(f)|/\|f\|_p\colon f\ne 0]$.

Suppose p and q are conjugate indices and $g\in L^q$. By Hölder's inequality,

$$\gamma_g(f)=\int fg\,d\mu \tag{19.8}$$

is defined for $f\in L^p$ and satisfies (19.7) if $M\ge\|g\|_q$; and γ_g is obviously linear. According to the *Riesz representation theorem*, this is the most general bounded linear functional in the case $p<\infty$:

Theorem 19.3. *Suppose that μ is σ-finite, that $1\le p<\infty$, and that q is conjugate to p. Every bounded linear functional on L^p has the form* (19.8) *for some $g\in L^q$; further,*

$$\|\gamma_g\|=\|g\|_q, \tag{19.9}$$

and g is unique up to a set of μ-measure 0.

†Part (ii) definitely requires $p<\infty$; see Problem 19.2.

The space of bounded linear functionals on L^p is called the *dual space*, or the *conjugate space*, and the theorem identifies L^q as the dual of L^p. Note that the theorem does not cover the case $p = \infty$.†

PROOF. *Case I: μ finite.* For A in $\mathscr{F}$, define $\varphi(A) = \gamma(I_A)$. The linearity of γ implies that φ is finitely additive. For the M of (19.7), $|\varphi(A)| \le M \cdot \|I_A\|_p = M \cdot |\mu(A)|^{1/p}$. If $A = \bigcup_n A_n$, where the A_n are disjoint, then $\varphi(A) = \sum_{n=1}^{N} \varphi(A_n) + \varphi(\bigcup_{n>N} A_n)$, and since $|\varphi(\bigcup_{n>N} A_n)| \le M\mu^{1/p}(\bigcup_{n>N} A_n) \to 0$, it follows that φ is an additive set function in the sense of (32.1).

The Jordan decomposition (32.2) represents φ as the difference of two finite measures φ^+ and φ^- with disjoint supports A^+ and A^-. If $\mu(A) = 0$, then $\varphi^+(A) = \varphi(A \cap A^+) \le M\mu^{1/p}(A) = 0$. Thus φ^+ is absolutely continuous with respect to μ and by the Radon-Nikodym theorem (p. 422) has an integrable density g^+: $\varphi^+(A) = \int_A g^+ \, d\mu$. Together with the same result for φ^-, this shows that there is an integrable g such that $\gamma(I_A) = \varphi(A) = \int_A g \, d\mu = \int I_A g \, d\mu$. Thus $\gamma(f) = \int fg \, d\mu$ for simple functions f in L^p.

Assume that this g lies in L^q, and define γ_g by the equation (19.8). Then γ and γ_g are bounded linear functionals that agree for simple functions; since the latter are dense (Theorem 19.2(i)), it follows by the continuity of γ and γ_g that they agree on all of L^p. It is therefore enough (in the case of finite μ) to prove $g \in L^q$. It will also be shown that $\|g\|_q$ is at most the M of (19.7); since $\|g\|_q$ does work as a bound in (19.7), (19.9) will follow. If $\gamma_g(f) \equiv 0$, (19.9) will imply that $g = 0$ almost everywhere, and for the general γ it will follow further that two functions g satisfying $\gamma_g(f) \equiv \gamma(f)$ must agree almost everywhere.

Assume that $1 < p, q < \infty$. Let g_n be simple functions such that $0 \le g_n \uparrow |g|^q$, and take $h_n = g_n^{1/p} \operatorname{sgn} g$. Then $h_n g = g_n^{1/p}|g| \ge g_n^{1/p} g_n^{1/q} = g_n$, and since h_n is simple, it follows that $\int g_n \, d\mu \le \int h_n g \, d\mu = \gamma_g(h_n) = \gamma(h_n) \le M \cdot \|h_n\|_p = M[\int g_n \, d\mu]^{1/p}$. Since $1 - 1/p = 1/q$, this gives $[\int g_n \, d\mu]^{1/q} \le M$. Now the monotone convergence theorem gives $g \in L^p$ and even $\|g\|_q \le M$.

Assume that $p = 1$ *and* $q = \infty$. In this case, $|\int fg \, d\mu| = |\gamma_g(f)| = |\gamma(f)| \le M \cdot \|f\|_1$ for simple functions f in L^1. Take $f = \operatorname{sgn} g \cdot I_{[|g| \ge \alpha]}$. Then $\alpha\mu[|g| \ge \alpha] \le \int I_{[|g| \ge \alpha]} \cdot |g| \, d\mu = \int fg \, d\mu \le M \cdot \|f\|_1 = M\mu[|g| \ge \alpha]$. If $\alpha > M$, this inequality gives $\mu[|g| \ge \alpha] = 0$; therefore $\|g\|_\infty = \|g\|_q \le M$ and $g \in L^\infty = L^q$.

Case II: μ σ-finite. Let A_n be sets such that $A_n \uparrow \Omega$ and $\mu(A_n) < \infty$. If $\mu_n(A) = \mu(A \cap A_n)$, then $|\gamma(fI_{A_n})| \le M \cdot \|fI_{A_n}\|_p = M \cdot [\int |f|^p d\mu_n]^{1/p}$ for $f \in L^p(fI_{A_n} \in L^p(\mu) \subset L^p(\mu_n))$. By the finite case, A_n supports a g_n in L^q such that $\gamma(fI_{A_n}) = \int fI_{A_n} g_n \, d\mu$ for $f \in L^p$, and $\|g_n\|_q \le M$. Because of uniqueness, g_{n+1} can be taken to agree with g_n on A_n $(L^p(\mu_{n+1}) \subset L^p(\mu_n))$. There is therefore a function g on Ω such that $g = g_n$ on A_n and $\|I_{A_n} g\|_q \le M$. It follows that $\|g\|_q \le M$ and $g \in L^q$. By the dominated convergence theorem and the continuity of γ, $f \in L^p$ implies $\int fg \, d\mu = \lim_n \int fI_{A_n} g \, d\mu = \lim_n \gamma(fI_{A_n}) = \gamma(f)$. Uniqueness follows as before. ∎

†Problem 19.3.

Weak Compactness

For $f \in L^p$ and $g \in L^q$, where p and q are conjugate, write

$$(f, g) = \int fg \, d\mu. \tag{19.10}$$

For fixed f in L^p, this defines a bounded linear functional on L^q; for fixed g in L^q, it defines a bounded linear functional on L^p. By Hölder's inequality,

$$|(f, g)| \le \|f\|_p \|g\|_q. \tag{19.11}$$

Suppose that f and f_n are elements of L^p. If $(f, g) = \lim_n (f_n, g)$ for each g in L^q, then f_n *converges weakly* to f. If $\|f - f_n\|_p \to 0$, then certainly f_n converges weakly to f, although the converse is false.[†]

The next theorem says in effect that if $p > 1$, then the unit ball $B_1^p = [f \in L^p: \|f\|_p \le 1]$ is compact in the topology of weak convergence.

Theorem 19.4. *Suppose that μ is σ-finite and $\mathcal{F}$ is countably generated. If $1 < p \le \infty$, then every sequence in B_1^p contains a subsequence converging weakly to an element of B_1^p.*

Suppose of elements f_n, f, and f' of L^p that f_n converges weakly both to f and to f'. Since, by hypothesis, μ is σ-finite and $p > 1$, Theorem 19.3 applies if the p and q there are interchanged. And now, since $(f, g) = (f', g)$ for all g in L^q, it follows by uniqueness that $f = f'$. Therefore, weak limits are unique under the present hypothesis. The assumption $p > 1$ is essential.[‡]

PROOF. Let q be conjugate to p $(1 \le q < \infty)$. By Theorem 19.2(ii), L^q contains a countable, dense set G. Add to G all finite, rational linear combinations of its elements; it is still countable. Suppose that $\{f_n\} \subset B_1^p$.

By (19.11), $|(f_n, g)| \le \|g\|_q$ for $g \in L^q$. Since $\{(f_n, g)\}$ is bounded, it is possible by the diagonal method [A14] to pass to a subsequence of $\{f_n\}$ along which, for each of the countably many g in G, the limit $\lim_n (f_n, g) = \gamma(g)$ exists and $|\gamma(g)| \le \|g\|_q$. For $g, g' \in G$, $|\gamma(g) - \gamma(g')| = \lim_n |(f_n, g - g')| \le \|g - g'\|_q$. Therefore, γ is uniformly continuous on G and so has a unique continuous extension to all of L^q. For $g, g' \in G$, $\gamma(g + g') = \lim_n (f_n, g + g') = \gamma(g) + \gamma(g')$; by continuity, this extends to all of L^q. For $g \in G$ and α rational, $\gamma(\alpha g) = \alpha \lim_n (f_n, g) = \alpha\gamma(g)$; by continuity, this extends to all real α and all g in L^q: γ is a linear functional on L^q. Finally, $|\gamma(g)| \le \|g\|_q$ extends from G to L^q by continuity, and γ is bounded in the sense of (19.7).

[†]Problem 19.4.
[‡]Problem 19.5.

By the Riesz representation theorem ($1 \le q < \infty$), there is an f in L^p (the space adjoint to L^q) such that $\gamma(g) = (f, g)$ for all g. Since γ has norm at most 1, (19.9) implies that $\|f\|_p \le 1$: f lies in B_1^p.

Now $(f, g) = \lim_n (f_n, g)$ for g in G. Suppose that g' is an arbitrary element of L^q, and choose g in G so that $\|g' - g\|_q < \epsilon$. Then

$$\begin{aligned}
&|(f, g') - (f_n, g')| \\
&\quad \le |(f, g') - (f, g)| + |(f, g) - (f_n, g)| + |(f_n, g) - (f_n, g')| \\
&\quad \le \|f\|_p \|g' - g\|_q + |(f, g) - (f_n, g)| + \|f_n\|_p \|g - g'\|_q \\
&\quad \le 2\epsilon + |(f, g) - (f_n, g)|.
\end{aligned}$$

Since $g \in G$, the last term here goes to 0, and hence $\lim_n (f_n, g') = (f, g')$ for all g' in L^q. Therefore, f_n converges weakly to f. ■

Some Decision Theory

The weak compactness of the unit ball in L^∞ has interesting implications for statistical decision theory. Suppose that μ is σ-finite and $\mathscr{F}$ is countably generated. Let $f_1, \ldots, f_k$ be probability densities with respect to μ—nonnegative and integrating to 1. Imagine that, for some i, ω is drawn from Ω according to the probability measure $P_i(A) = \int_A f_i \, d\mu$. The statistical problem is to decide, on the basis of an observed ω, which f_i is the right one.

Assume that if the right density is f_i, then a statistician choosing f_j incurs a nonnegative *loss* $L(j|i)$. A *decision rule* is a vector function $\delta(\omega) = (\delta_1(\omega), \ldots, \delta_k(\omega))$, where the $\delta_i(\omega)$ are nonnegative and add to 1: the statistician, observing ω, selects f_i with probability $\delta_i(\omega)$. If, for each ω, $\delta_i(\omega)$ is 1 for one i and 0 for the others, δ is a *nonrandomized* rule; otherwise, it is a *randomized* rule. Let D be the set of all rules. The problem is to choose, in some more or less rational way that connects up with the losses $L(j|i)$, a rule δ from D.

The *risk* corresponding to δ and f_i is

$$R_i(\delta) = \int \left[\sum_j \delta_j(\omega) L(j|i) \right] f_i(\omega) \mu(d\omega),$$

which can be interpreted as the loss a statistician using δ can expect if f_i is the right density. The *risk point* for δ is $R(\delta) = (R_1(\delta), \ldots, R_k(\delta))$. If $R_i(\delta') \le R_i(\delta)$ for all i and $R_i(\delta') < R_i(\delta)$ for some i—that is, if the point $R(\delta')$ is "southwest" of $R(\delta)$—then of course δ' is taken as being *better* than δ. There is in general no rule better than all the others. (Different rules can have the same risk point, but they are then indistinguishable as regards the decision problem.)

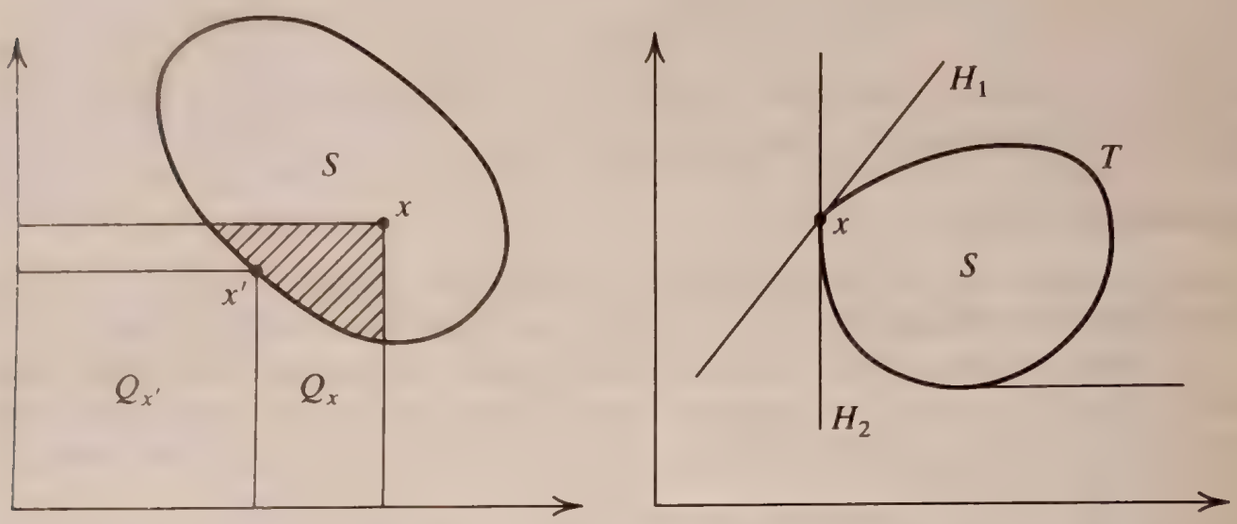
The *risk set* is the collection S of all the risk points; it is a bounded set in the first orthant of R^k. To avoid trivialities, assume that S does not contain the origin (as would happen for example if the $L(j|i)$ were all 0).

Suppose that δ and δ' are elements of D, and λ and λ' are nonnegative and add to 1. If $\delta_i''(\omega) = \lambda \delta_i(\omega) + \lambda' \delta_i'(\omega)$, then δ'' is in D and $R(\delta'') = \lambda R(\delta) + \lambda' R(\delta')$. Therefore, S is a convex set.

Lying much deeper is the fact that S is compact. Given points $x^{(n)}$ in S, choose rules $\delta^{(n)}$ such that $R(\delta^{(n)}) = x^{(n)}$. Now $\delta_j^{(n)}(\cdot)$ is an element of L^∞, in fact of B_1^∞, and so by Theorem 19.4 there is a subsequence along which, for each $j = 1, \ldots, k$, $\delta_j^{(n)}$

converges weakly to a function δ_j in B_1^∞. If $\mu(A) < \infty$, then $\int \delta_j I_A \, d\mu = \lim_n \int \delta_j^{(n)} I_A \, d\mu \geq 0$ and $\int (1 - \sum_j \delta_j) I_A \, d\mu = \lim_n \int (1 - \sum_j \delta_j^{(n)}) I_A \, d\mu = 0$, so that $\delta_j \geq 0$ and $\sum_j \delta_j = 1$ almost everywhere on A. Since μ is σ-finite, the δ_j can be altered on a set of μ-measure 0 in such a way as to ensure that $\delta = (\delta_1, \ldots, \delta_k)$ is an element of D. But, along the subsequence, $x^{(n)} \to R(\delta)$. Therefore: *The risk set is compact and convex.*

The rest is geometry. For x in R^k, let Q_x be the set of x' such that $0 \leq x_i' \leq x_i$ for all i. If $x = R(\delta)$ and $x' = R(\delta')$, then δ' is better than δ if and only if $x' \in Q_x$ and $x' \neq x$. A rule δ is *admissible* if there exists no δ' better than δ; it makes no sense to use a rule that is not admissible. Geometrically, admissibility means that, for $x = R(\delta)$, $S \cap Q_x$ consists of x alone.



Let $x = R(\delta)$ be given, and suppose that δ is not admissible. Since $S \cap Q_x$ is compact, it contains a point x' nearest the origin (x' unique, since $S \cap Q_x$ is convex as well as compact); let δ' be a corresponding rule: $x' = R(\delta')$. Since δ is not admissible, $x' \neq x$, and δ' is better than δ. If $S \cap Q_{x'}$ contained a point distinct from x', it would be a point of $S \cap Q_x$ nearer the origin than x', which is impossible. This means that $Q_{x'}$ contains no point of S other than x' itself, which means in turn that δ' is admissible. Therefore, if δ is not itself admissible, there is a δ' that is admissible and is better than δ. This is expressed by saying that the class of admissible rules is *complete.*

Let $p = (p_1, \ldots, p_k)$ be a probability vector, and view p_i as an a priori probability that f_i is the correct density. A rule δ has *Bayes risk* $R(p, \delta) = \sum_i p_i R_i(\delta)$ with respect to p. This is a kind of compound risk: f_i is correct with probability p_i, and the statistician chooses f_j with probability $\delta_j(\omega)$. A *Bayes rule* is one that minimizes the Bayes risk for a given p. In this case, take $\alpha = R(p, \delta)$ and consider the hyperplane

$$H = \left[z : \sum_i p_i z_i = \alpha \right] \tag{19.12}$$

and the half space

$$H^+ = \left[z : \sum_i p_i z_i \geq \alpha \right]. \tag{19.13}$$

Then $x = R(\delta)$ lies on H, and S is contained in H^+: x is on the boundary of S, and

H is a supporting hyperplane. If $p_i > 0$ for all i, then Q_x meets S only at x, and so δ is admissible.

Suppose now that δ is admissible, so that $x = R(\delta)$ is the only point in $S \cap Q_x$ and x lies on the boundary of S. The problem is to show that δ is a Bayes rule, which means finding a supporting hyperplane (19.12) corresponding to a probability vector p. Let T consist of those y for which Q_y meets S. Then T is convex: given a convex combination $y'' = \lambda y + \lambda' y'$ of points in T, choose in S points z and z' southwest of y and y', respectively, and note that $z'' = \lambda z + \lambda' z'$ lies in S and is southwest of y''. Since S meets Q_x only in the point x, the same is true of T, so that x is a boundary point of T as well as of S. Let (19.12) ($p \neq 0$) be a supporting hyperplane through x: $x \in H$ and $T \subset H^+$. If $p_{i_0} < 0$, take $z_{i_0} = x_{i_0} + 1$ and take $z_i = x_i$ for the other i; then z lies in T but not in H^+, a contradiction. (The right-hand figure shows the role of T: the planes H_1 and H_2 both support S, but only H_2 supports T and only H_2 corresponds to a probability vector.) Thus $p_i \geq 0$ for all i, and since $\sum_i p_i = 1$ can be arranged by normalization, δ is indeed a Bayes rule. Therefore: *The admissible rules are Bayes rules, and they form a complete class.*

The Space L^2

The space L^2 is special because $p = 2$ is its own conjugate index. If $f, g \in L^2$, the *inner product* $(f, g) = \int fg\, d\mu$ is well defined, and by (19.11)—write $\|f\|$ in place of $\|f\|_2$—$|(f, g)| \leq \|f\| \cdot \|g\|$. This is the *Schwarz* (or Cauchy–Schwarz) inequality. If one of f and g is fixed, (f, g) is a bounded (hence continuous) linear functional in the other. Further, $(f, g) = (g, f)$, the norm is given by $\|f\|^2 = (f, f)$, and L^2 is complete under the metric $d(f, g) = \|f - g\|$. A *Hilbert space* is a vector space on which is defined an inner product having all these properties.

The Hilbert space L^2 is quite like Euclidean space. If $(f, g) = 0$, then f and g are *orthogonal*, and orthogonality is like perpendicularity. If $f_1, \ldots, f_n$ are orthogonal (in pairs), then by linearity, $(\sum_i f_i, \sum_j f_j) = \sum_i \sum_j (f_i, f_j) = \sum_i (f_i, f_i)$: $\|\sum_i f_i\|^2 = \sum_i \|f_i\|^2$. This is a version of the Pythagorean theorem. If f and g are orthogonal, write $f \perp g$. For every f, $f \perp 0$.

Suppose now that μ is σ-finite and $\mathscr{F}$ is countably generated, so that L^2 is separable as a metric space. The construction that follows gives a sequence (finite or infinite) $\varphi_1, \varphi_2, \ldots$ that is *orthonormal* in the sense that $\|\varphi_n\| = 1$ for all n and $(\varphi_m, \varphi_n) = 0$ for $m \neq n$, and is *complete* in the sense that $(f, \varphi_n) = 0$ for all n implies $f = 0$—so that the orthonormal system cannot be enlarged. Start with a sequence $f_1, f_2, \ldots$ that is dense in L^2. Define $g_1, g_2, \ldots$ inductively: Let $g_1 = f_1$. Suppose that $g_1, \ldots, g_n$ have been defined and are orthogonal. Define $g_{n+1} = f_{n+1} - \sum_{i=1}^n \alpha_{ni} g_i$, where α_{ni} is $(f_{n+1}, g_i)/\|g_i\|^2$ if $g_i \neq 0$ and is arbitrary if $g_i = 0$. Then g_{n+1} is orthogonal to $g_1, \ldots, g_n$, and f_{n+1} is a linear combination of $g_1, \ldots, g_{n+1}$. This, the Gram–Schmidt method, gives an orthogonal sequence $g_1, g_2, \ldots$ with the property that the finite linear combinations of the g_n include all the f_n and are therefore dense in L^2. If $g_n \neq 0$, take $\varphi_n = g_n/\|g_n\|$; if $g_n = 0$, discard it from the sequence. Then $\varphi_1, \varphi_2, \ldots$ is orthonormal, and the finite linear combinations of the φ_n are still dense. It can happen that all but finitely many of the g_n are 0, in which

case there are only finitely many of the φ_n. In what follows it is assumed that $\varphi_1, \varphi_2, \ldots$ is an infinite sequence; the finite case is analogous and somewhat simpler.

Suppose that f is orthogonal to all the φ_n. If a_i are arbitrary scalars, then $f, a_1\varphi_1, \ldots, a_n\varphi_n$ is an orthogonal set, and by the Pythagorean property, $\|f - \sum_{i=1}^n a_i\varphi_i\|^2 = \|f\|^2 + \sum_{i=1}^n a_i^2 \geq \|f\|^2$. If $\|f\| > 0$, then f cannot be approximated by finite linear combinations of the φ_n, a contradiction: $\varphi_1, \varphi_2, \ldots$ *is a complete orthonormal system*.

Consider now a sequence $a_1, a_2, \ldots$ of scalars for which $\sum_{i=1}^\infty a_i^2$ converges. If $s_n = \sum_{i=1}^n a_i\varphi_i$, then the Pythagorean theorem gives $\|s_n - s_m\|^2 = \sum_{m<i\leq n} a_i^2$. Since the scalar series converges, $\{s_n\}$ is fundamental and therefore by Theorem 19.1 converges to some g in L^2. Thus $g = \lim_n \sum_{i=1}^n a_i\varphi_i$, which it is natural to express as $g = \sum_{i=1}^\infty a_i\varphi_i$. The series (that is to say, the sequence of partial sums) converges to g in the mean of order 2 (not almost everywhere). By the following argument, every element of L^2 has a unique representation in this form.

The *Fourier coefficients* of f with respect to $\{\varphi_i\}$ are the inner products $a_i = (f, \varphi_i)$. For each n, $0 \leq \|f - \sum_{i=1}^n a_i\varphi_i\|^2 = \|f\|^2 - 2\sum_i a_i(f, \varphi_i) + \sum_{ij} a_i a_j(\varphi_i, \varphi_j) = \|f\|^2 - \sum_{i=1}^n a_i^2$, and hence, n being arbitrary, $\sum_{i=1}^\infty a_i^2 \leq \|f\|^2$. By the argument above, the series $\sum_{i=1}^\infty a_i\varphi_i$ therefore converges. By linearity, $(f - \sum_{i=1}^n a_i\varphi_i, \varphi_j) = 0$ for $n \geq j$, and by continuity, $(f - \sum_{i=1}^\infty a_i\varphi_i, \varphi_j) = 0$. Therefore, $f - \sum_{i=1}^\infty a_i\varphi_i$ is orthogonal to each φ_j and by completeness must be 0:

$$f = \sum_{i=1}^{\infty} (f, \varphi_i)\varphi_i. \tag{19.14}$$

This is the *Fourier representation* of f. It is unique because if $f = \sum_{i=1}^\infty a_i\varphi_i$ is 0 ($\sum a_i^2 < \infty$), then $a_j = (f, \varphi_j) = 0$. Because of (19.14), $\{\varphi_n\}$ is also called an orthonormal *basis* for L^2.

A subset M of L^2 is a *subspace* if it is closed both algebraically ($f, f' \in M$ implies $\alpha f + \alpha' f' \in M$) and topologically ($f_n \in M$, $f_n \to f$ implies $f \in M$). If L^2 is separable, then so is the subspace M, and the construction above carries over: M contains an orthonormal system $\{\varphi_n\}$ that is *complete in* M, in the sense that $f = 0$ if $(f, \varphi_n) = 0$ for all n *and if* $f \in M$. And each f in M has the unique Fourier representation (19.14). Even if f does not lie in M, $\sum_{i=1}^\infty (f, \varphi_i)^2$ converges, so that $\sum_{i=1}^\infty (f, \varphi_i)\varphi_i$ is well defined.

This leads to a powerful idea, that of *orthogonal projection* onto M. For an orthonormal basis $\{\varphi_i\}$ of M, define $P_M f = \sum_{i=1}^\infty (f, \varphi_i)\varphi_i$ for all f in L^2 (not just for f in M). Clearly, $P_M f \in M$. Further, $f - \sum_{i=1}^n (f, \varphi_i)\varphi_i \perp \varphi_j$ for $n \geq j$ by linearity, so that $f - P_M f \perp \varphi_j$ by continuity. But if $f - P_M f$ is orthogonal to each φ_j, then, again by linearity and continuity, it is orthogonal to the general element $\sum_{j=1}^\infty b_j\varphi_j$ of M. Therefore, $P_M f \in M$ and $f - P_M f \perp M$. The map $f \to P_M f$ is the orthogonal projection on M.

The fundamental properties of P_M are these:

(i) $g \in M$ and $f - g \perp M$ together imply $g = P_M f$;
(ii) $f \in M$ implies $P_M f = f$;
(iii) $g \in M$ implies $\|f - g\| \geq \|f - P_M f\|$;
(iv) $P_M(\alpha f + \alpha' f') = \alpha P_M f + \alpha' P_M f'$.

Property (i) says that $P_M f$ is uniquely determined by the two conditions $P_M f \in M$ and $f - P_M \perp M$. To prove it, suppose that $g, g' \in M$, $f - g \perp M$, and $f - g' \perp M$. Then $g - g' \in M$ and $g - g' \perp M$, so that $g - g'$ is orthogonal to itself and hence $\|g - g'\|^2 = 0$: $g = g'$. Thus the mapping P_M is independent of the particular basis $\{\varphi_i\}$; it is determined by M alone.

Clearly, (ii) follows from (i); it implies that P_M is idempotent in the sense that $P_M^2 f = P_M f$. As for (iii), if g lies in M, so does $P_M f - g$, so that, by the Pythagorean relation, $\|f - g\|^2 = \|f - P_M f\|^2 + \|P_M f - g\|^2 \geq \|f - P_M f\|^2$; the inequality is strict if $g \neq P_M f$. Thus $P_M f$ is the unique point of M lying nearest to f. Property (iv), linearity, follows from (i).

An Estimation Problem

First, the technical setting: Let $(\Omega, \mathcal{F}, \mu)$ and $(\Theta, \mathcal{E}, \pi)$ be a σ-finite space and a probability space, and assume that $\mathcal{F}$ and $\mathcal{E}$ are countably generated. Let $f_\theta(\omega)$ be a nonnegative function on $\Theta \times \Omega$, measurable $\mathcal{E} \times \mathcal{F}$, and assume that $\int_\Omega f_\theta(\omega)\mu(d\omega) = 1$ for each $\theta \in \Theta$. For some unknown value of θ, ω is drawn from Ω according to the probabilities $P_\theta(A) = \int_A f_\theta(\omega)\mu(d\omega)$, and the statistical problem is to estimate the value of $g(\theta)$, where g is a real function on Θ. The statistician knows the functions $f_\cdot(\cdot)$ and $g(\cdot)$, as well as the value of ω; it is the value of θ that is unknown.

For an example, take Ω to be the line, $f(\omega)$ a function known to the statistician, and $f_\theta(\omega) = \alpha f(\alpha\omega + \beta)$, where $\theta = (\alpha, \beta)$ specifies unknown scale and location parameters; the problem is to estimate $g(\theta) = \alpha$, say. Or, more simply, as in the exponential case (14.7), take $f_\theta(\omega) = \alpha f(\alpha\omega)$, where $\theta = g(\theta) = \alpha$.

An *estimator* of $g(\theta)$ is a function $t(\omega)$. It is *unbiased* if

$$\int_\Omega t(\omega) f_\theta(\omega)\mu(d\omega) = g(\theta) \tag{19.15}$$

for all θ in Θ (assume the integral exists); this condition means that the estimate is on target in an average sense. A natural loss function is $(t(\omega) - g(\theta))^2$, and if f_θ is the correct density, the *risk* is taken to be $\int_\Omega (t(\omega) - g(\theta))^2 f_\theta(\omega)\mu(d\theta)$.

If the probability measure π is regarded as an a priori distribution for the unknown θ, the *Bayes risk* of t is

$$R(\pi, t) = \int_\Theta \int_\Omega \big(t(\omega) - g(\theta)\big)^2 f_\theta(\omega)\mu(d\omega)\pi(d\theta); \tag{19.16}$$

this integral, assumed finite, can be viewed as a joint integral or as an iterated integral (Fubini's theorem). And now t_0 is a *Bayes estimator* of g with respect to π if it minimizes $R(\pi, t)$ over t. This is analogous to the Bayes rules discussed earlier. The

following simple projection argument shows that, except in trivial cases, no Bayes estimator is unbaised.†

Let Q be the probability measure on $\mathscr{E}\times\mathscr{F}$ having density $f_\theta(\omega)$ with respect to $\pi\times\mu$, and let L^2 be the space of square-integrable functions on $(\Theta\times\Omega, \mathscr{E}\times\mathscr{F}, Q)$. Then Q is finite and $\mathscr{E}\times\mathscr{F}$ is countably generated. Recall that an element of L^2 is an equivalence class of functions that are equal almost everywhere with respect to Q. Let G be the class of elements of L^2 containing a function of the form $\bar{g}(\theta,\omega)=g(\omega)$—functions of θ alone. Then G is a subspace. (That G is algebraically closed is clear; if $f_n\in G$ and $\|f_n-f\|\to 0$, then—see the proof of Theorem 19.1—some subsequence converges to f outside a set of Q-measure 0, and it follows easily that $f\in G$.) Similarly, let T be the subspace of functions of ω alone: $\bar{t}(\theta,\omega)=t(\omega)$. Consider only functions g and their estimators t for which the corresponding $\bar{g}$ and $\bar{t}$ are in L^2.

Suppose now that t_0 is both an unbiased estimator of g_0 and a Bayes estimator of g_0 with respect to π. By (19.16) for g_0, $R(\pi,t)=\|\bar{t}-\bar{g}_0\|^2$, and since t_0 is a Bayes estimator of g_0, it follows that $\|\bar{t}_0-\bar{g}_0\|^2\le\|\bar{t}-\bar{g}_0\|^2$ for all $\bar{t}$ in T. This means that $\bar{t}_0$ is the orthogonal projection of $\bar{g}_0$ on the subspace T and hence that $\bar{g}_0-\bar{t}_0\perp\bar{t}_0$. On the other hand, from the assumption that t_0 is an unbiased estimator of g_0, it follows that, for every $\bar{g}(\theta,\omega)=g(\theta)$ in G,

$$(\bar{t}_0-\bar{g}_0,\bar{g})=\int_\Theta\int_\Omega(t_0(\omega)-g_0(\theta))g(\theta)f_\theta(\omega)\mu(d\omega)\pi(d\theta)$$

$$=\int_\Theta g(\theta)\left[\int_\Omega(t_0(\omega)-g_0(\theta))f_\theta(\omega)\mu(d\omega)\right]\pi(d\theta)=0.$$

This means that $\bar{t}_0-\bar{g}_0\perp G$: $\bar{g}_0$ is the orthogonal projection of $\bar{t}_0$ on the subspace G. But $\bar{g}_0-\bar{t}_0\perp\bar{t}_0$ and $\bar{t}_0-\bar{g}_0\perp\bar{g}_0$ together imply that $\bar{t}_0-\bar{g}_0$ is orthogonal to itself: $\bar{t}_0=\bar{g}_0$. Therefore, $t_0(\omega)=\bar{t}_0(\theta,\omega)=\bar{g}_0(\theta,\omega)=g_0(\theta)$ for (θ,ω) outside a set of Q-measure 0.

This implies that t_0 and g_0 are essentially constant. Suppose for simplicity that $f_\theta(\omega)>0$ for all (θ,ω), so that (Theorem 15.2) $(\pi\times\mu)[(\theta,\omega)\colon t_0(\omega)\ne g_0(\theta)]=0$. By Fubini's theorem, there is a θ such that, if $a=g_0(\theta)$, then $\mu[\omega\colon t_0(\omega)\ne a]=0$; and there is an ω such that, if $b=t_0(\omega)$, then $\pi[\theta\colon g_0(\theta)\ne b]=0$. It follows that, for (θ,ω) outside a set of $(\pi\times\mu)$-measure 0, $t_0(\omega)$ and $g_0(\theta)$ have the common value $a=b$: $\pi[\theta\colon g_0(\theta)=a]=1$ and $P_\theta[\omega\colon t_0(\omega)=a]=1$ for all θ in Θ.

PROBLEMS

19.1. Suppose that $\mu(\Omega)<\infty$ and $f\in L^\infty$. Show that $\|f\|_p\uparrow\|f\|_\infty$.

19.2. (a) Show that $L^\infty((0,1],\mathscr{B},\lambda)$ is not separable.

(b) Show that $L^p((0,1],\mathscr{B},\mu)$ is not separable if μ is counting measure (μ is not σ-finite).

(c) Show that $L^p(\Omega,\mathscr{F},P)$ is not separable if (Theorem 36.2) there is on the space an independent stochastic process $[X_t\colon 0\le t\le 1]$ such that X_t takes the values ±1 with probability $\frac{1}{2}$ each ($\mathscr{F}$ is not countably generated).

†This is interesting because of the close connection between Bayes rules and admissibility; see BERGER, pp. 546 ff.

19.3. Show that Theorem 19.3 fails for $L^\infty((0,1], \mathscr{B}, \lambda)$. *Hint:* Take $\gamma(f)$ to be a Banach limit of $n\int_0^{1/n} f(x)\,dx$.

19.4. Consider weak convergence in $L^p((0,1], \mathscr{B}, \lambda)$.

(a) For the case $p = \infty$, find functions f_n and f such that f_n goes weakly to f but $\|f - f_n\|_p$ does not go to 0.

(b) Do the same for $p = 2$.

19.5. Show that the unit ball in $L^1((0,1], \mathscr{B}, \lambda)$ is not weakly compact.

19.6. Show that a Bayes rule corresponding to $p = (p_1, \ldots, p_k)$ may not be admissible if $p_i = 0$ for some i. But there will be a better Bayes rule that *is* admissible.

19.7. *The Neyman–Pearson lemma.* Suppose f_1 and f_2 are rival densities and $L(j|i)$ is 0 or 1 as $j = i$ or $j \neq i$, so that $R_i(\delta)$ is the probability of choosing the opposite density when f_i is the right one. Suppose of δ that $\delta_2(\omega) = 1$ if $f_2(\omega) > tf_1(\omega)$ and $\delta_2(\omega) = 0$ if $f_2(\omega) < tf_1(\omega)$, where $t > 0$. Show that δ is admissible: For any rule δ', $\int \delta_2' f_1\,d\mu < \int \delta_2 f_1\,d\mu$ implies $\int \delta_1' f_2\,d\mu > \int \delta_1 f_2\,d\mu$. *Hint:* $\int (\delta_2 - \delta_2')(f_2 - tf_1)\,d\mu \geq 0$, since the integrand is nonnegative.

19.8. The classical orthonormal basis for $L^2[0, 2\pi]$ with Lebesgue measure is the trigonometric system

$$(19.17)\quad (2\pi)^{-1}, \qquad \pi^{-1/2}\sin nx, \qquad \pi^{-1/2}\cos nx, \qquad n = 1, 2, \ldots.$$

Prove orthonormality. *Hint:* Express the sines and cosines in terms of $e^{inx} \pm e^{-inx}$, multiply out the products, and use the fact that $\int_0^{2\pi} e^{imx}\,dx$ is 2π or 0 as $m = 0$ or $m \neq 0$. (For the completeness of the trigonometric system, see Problem 26.26.)

19.9. Drop the assumption that L^2 is separable. Order by inclusion the orthonormal systems in L^2, and let (Zorn's lemma) $\Phi = [\varphi_\gamma\colon \gamma \in \Gamma]$ be maximal.

(a) Show that $\Gamma_f = [\gamma\colon (f, \varphi_\gamma) \neq 0]$ is countable. *Hint:* Use $\sum_{i=1}^n (f, \varphi_{\gamma_i})^2 \leq \|f\|^2$ and the argument for Theorem 10.2(iv).

(b) Let $Pf = \sum_{\gamma \in \Gamma_f} (f, \varphi_\gamma)\varphi_\gamma$. Show that $f - Pf \perp \Phi$ and hence (maximality) $f = Pf$. Thus Φ is an orthonormal basis.

(c) Show that Φ is countable if and only if L^2 is separable.

(d) Now take Φ to be a maximal orthonormal system in a subspace M, and define $P_M f = \sum_{\gamma \in \Gamma_f} (f, \varphi_\gamma)\varphi_\gamma$. Show that $P_M f \in M$ and $f - P_M f \perp \Phi$, that $g = P_M g$ if $g \in M$, and that $f - P_M f \perp M$. This defines the general orthogonal projection.

CHAPTER 4

Random Variables and Expected Values

SECTION 20. RANDOM VARIABLES AND DISTRIBUTIONS

This section and the next cover random variables and the machinery for dealing with them—expected values, distributions, moment generating functions, independence, convolution.

Random Variables and Vectors

A *random variable* on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued function $X = X(\omega)$ measurable $\mathcal{F}$. Sections 5 through 9 dealt with random variables of a special kind, namely simple random variables, those with finite range. All concepts and facts concerning real measurable functions carry over to random variables; any changes are matters of viewpoint, notation, and terminology only.

The positive and negative parts X^+ and X^- of X are defined as in (15.4) and (15.5). Theorem 13.5 also applies: Define

$$\psi_n(x) = \begin{cases} (k-1)2^{-n} & \text{if } (k-1)2^{-n} \le x < k2^{-n}, \\ & \qquad 1 \le k \le n2^n, \\ n & \text{if } x \ge n. \end{cases} \tag{20.1}$$

If X is nonnegative and $X_n = \psi_n(X)$, then $0 \le X_n \uparrow X$. If X is not necessarily nonnegative, define

$$X_n = \begin{cases} \psi_n(X) & \text{if } X \ge 0, \\ -\psi_n(-X) & \text{if } X \le 0. \end{cases} \tag{20.2}$$

(This is the same as (13.6).) Then $0 \le X_n(\omega) \uparrow X(\omega)$ if $X(\omega) \ge 0$ and $0 \ge$

$X_n(\omega)\downarrow X(\omega)$ if $X(\omega)\le 0$; and $|X_n(\omega)|\uparrow|X(\omega)|$ for every ω. The random variable X_n is in each case simple.

A *random vector* is a mapping from Ω to R^k that is measurable $\mathscr{F}$. Any mapping from Ω to R^k must have the form $\omega\to X(\omega)=(X_1(\omega),\ldots,X_k(\omega))$, where each $X_i(\omega)$ is real; as shown in Section 13 (see (13.2)), X is measurable if and only if each X_i is. Thus a random vector is simply a k-tuple $X=(X_1,\ldots,X_k)$ of random variables.

Subfields

If $\mathscr{G}$ is a σ-field for which $\mathscr{G}\subset\mathscr{F}$, a k-dimensional random vector X is of course measurable $\mathscr{G}$ if $[\omega\colon X(\omega)\in H]\in\mathscr{G}$ for every H in $\mathscr{R}^k$. The σ-field $\sigma(X)$ generated by X is the smallest σ-field with respect to which it is measurable. The σ-field generated by a collection of random vectors is the smallest σ-field with respect to which each one is measurable.

As explained in Sections 4 and 5, a sub-σ-field corresponds to partial information about ω. The information contained in $\sigma(X)=\sigma(X_1,\ldots,X_k)$ consists of the k numbers $X_1(\omega),\ldots,X_k(\omega)$.[†] The following theorem is the analogue of Theorem 5.1, but there are technical complications in its proof.

Theorem 20.1. *Let $X=(X_1,\ldots,X_k)$ be a random vector.*

(i) *The σ-field $\sigma(X)=\sigma(X_1,\ldots,X_k)$ consists exactly of the sets $[X\in H]$ for $H\in\mathscr{R}^k$.*

(ii) *In order that a random variable Y be measurable $\sigma(X)=\sigma(X_1,\ldots,X_k)$ it is necessary and sufficient that there exist a measurable map $f\colon R^k\to R^1$ such that $Y(\omega)=f(X_1(\omega),\ldots,X_k(\omega))$ for all ω.*

PROOF. The class $\mathscr{G}$ of sets of the form $[X\in H]$ for $H\in\mathscr{R}^k$ is a σ-field. Since X is measurable $\sigma(X)$, $\mathscr{G}\subset\sigma(X)$. Since X is measurable $\mathscr{G}$, $\sigma(X)\subset\mathscr{G}$. Hence part (i).

Measurability of f in part (ii) refers of course to measurability $\mathscr{R}^k/\mathscr{R}^1$. The sufficiency is easy: if such an f exists, Theorem 13.1(ii) implies that Y is measurable $\sigma(X)$.

To prove necessity,[‡] suppose at first that Y is a simple random variable, and let $y_1,\ldots,y_m$ be its different possible values. Since $A_i=[\omega\colon Y(\omega)=y_i]$ lies in $\sigma(X)$, it must by part (i) have the form $[\omega\colon X(\omega)\in H_i]$ for some H_i in $\mathscr{R}^k$. Put $f=\sum_i y_i I_{H_i}$; certainly f is measurable. Since the A_i are disjoint, no $X(\omega)$ can lie in more than one H_i (even though the latter need not be disjoint), and hence $f(X(\omega))=Y(\omega)$.

[†]The partition defined by (4.16) consists of the sets $[\omega\colon X(\omega)=x]$ for $x\in R^k$.

[‡]For a general version of this argument, see Problem 13.3.

To treat the general case, consider simple random variables Y_n such that $Y_n(\omega) \to Y(\omega)$ for each ω. For each n, there is a measurable function f_n: $R^k \to R^1$ such that $Y_n(\omega) = f_n(X(\omega))$ for all ω. Let M be the set of x in R^k for which $\{f_n(x)\}$ converges; by Theorem 13.4(iii), M lies in $\mathscr{R}^k$. Let $f(x) = \lim_n f_n(x)$ for x in M, and let $f(x) = 0$ for x in $R^k - M$. Since $f = \lim_n f_n I_M$ and $f_n I_M$ is measurable, f is measurable by Theorem 13.4(ii). For each ω, $Y(\omega) = \lim_n f_n(X(\omega))$; this implies in the first place that $X(\omega)$ lies in M and in the second place that $Y(\omega) = \lim_n f_n(X(\omega)) = f(X(\omega))$. ■

Distributions

The distribution or law of a random variable X was in Section 14 defined as the probability measure on the line given by $\mu = PX^{-1}$ (see (13.7)), or

$$\mu(A) = P[X \in A], \qquad A \in \mathscr{R}^1. \tag{20.3}$$

The distribution function of X was defined by

$$F(x) = \mu(-\infty, x] = P[X \le x] \tag{20.4}$$

for real x. The left-hand limit of F satisfies

$$\begin{aligned} F(x-) &= \mu(-\infty, x) = P[X < x], \\ F(x) - F(x-) &= \mu\{x\} = P[X = x], \end{aligned} \tag{20.5}$$

and F has at most countably many discontinuities. Further, F is nondecreasing and right-continuous, and $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$. By Theorem 14.1, for each F with these properties there exists on some probability space a random variable having F as its distribution function.

A support for μ is a Borel set S for which $\mu(S) = 1$. A random variable, its distribution, and its distribution function are *discrete* if μ has a countable support $S = \{x_1, x_2, \ldots\}$. In this case μ is completely determined by the values $\mu\{x_1\}, \mu\{x_2\}, \ldots$.

A familiar discrete distribution is the *binomial*:

$$P[X = r] = \mu\{r\} = \binom{n}{r} p^r (1-p)^{n-r}, \qquad r = 0, 1, \ldots, n. \tag{20.6}$$

There are many random variables, on many spaces, with this distribution: If $\{X_k\}$ is an independent sequence such that $P[X_k = 1] = p$ and $P[X_k = 0] = 1 - p$ (see Theorem 5.3), then X could be $\sum_{i=1}^{n} X_i$, or $\sum_{i=9}^{8+n} X_i$, or the sum of any n of the X_i, Or Ω could be $\{0, 1, \ldots, n\}$ if $\mathscr{F}$ consists of all subsets, $P\{r\} = \mu\{r\}$, $r = 0, 1, \ldots, n$, and $X(r) \equiv r$. Or again the space and random variable could be those given by the construction in either of the two proofs of Theorem 14.1. These examples show that, although the distribution of a

random variable X contains all the information about the probabilistic behavior of X itself, it contains beyond this no further information about the underlying probability space $(\Omega, \mathscr{F}, P)$ or about the interaction of X with other random variables on the space.

Another common discrete distribution is the *Poisson* distribution with parameter $\lambda > 0$:

$$P[X = r] = \mu\{r\} = e^{-\lambda}\frac{\lambda^r}{r!}, \qquad r = 0, 1, \ldots. \tag{20.7}$$

A *constant* c can be regarded as a discrete random variable with $X(\omega) \equiv c$. In this case $P[X = c] = \mu\{c\} = 1$. For an artificial discrete example, let $\{x_1, x_2, \ldots\}$ be an enumeration of the rationals, and put

$$\mu\{x_r\} = 2^{-r}; \tag{20.8}$$

the point of the example is that the support need not be contained in a lattice.

A random variable and its distribution have *density* f with respect to Lebesgue measure if f is a nonnegative Borel function on R^1 and

$$P[X \in A] = \mu(A) = \int_A f(x)\,dx, \qquad A \in \mathscr{R}^1. \tag{20.9}$$

In other words, the requirement is that μ have density f with respect to Lebesgue measure λ in the sense of (16.11). The density is assumed to be with respect to λ if no other measure is specified.

Taking $A = R^1$ in (20.9) shows that f must integrate to 1. Note that f is determined only to within a set of Lebesgue measure 0: if $f = g$ except on a set of Lebesgue measure 0, then g can also serve as a density for X and μ.

It follows by Theorem 3.3 that (20.9) holds for every Borel set A if it holds for every interval—that is, if

$$F(b) - F(a) = \int_a^b f(x)\,dx$$

holds for every a and b. Note that F need not differentiate to f everywhere (see (20.13), for example); all that is required is that f integrate properly—that (20.9) hold. On the other hand, if F does differentiate to f and f is continuous, it follows by the fundamental theorem of calculus that f is indeed a density for F.[†]

[†]The general question of the relation between differentiation and integration is taken up in Section 31.

For the *exponential distribution* with parameter $\alpha > 0$, the density is

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ \alpha e^{-\alpha x} & \text{if } x \geq 0. \end{cases} \tag{20.10}$$

The corresponding distribution function

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\alpha x} & \text{if } \geq 0 \end{cases} \tag{20.11}$$

was studied in Section 14.

For the *normal distribution* with parameters m and σ, $\sigma > 0$,

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2} \right], \qquad -\infty < x < \infty; \tag{20.12}$$

a change of variable together with (18.10) shows that f does integrate to 1. For the *standard* normal distribution, $m = 0$ and $\sigma = 1$.

For the *uniform* distribution over an interval $(a, b]$,

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x \leq b, \\ 0 & \text{otherwise.} \end{cases} \tag{20.13}$$

The distribution function F is useful if it has a simple expression, as in (20.11). It is ordinarily simpler to describe μ by means of a density $f(x)$ or discrete probabilities $\mu\{x_r\}$.

If F comes from a density, it is continuous. In the discrete case, F increases in jumps; the example (20.8), in which the points of discontinuity are dense, shows that it may nonetheless be very irregular. There exist distributions that are not discrete but are not continuous either. An example is $\mu(A) = \frac{1}{2}\mu_1(A) + \frac{1}{2}\mu_2(A)$ for μ_1 discrete and μ_2 coming from a density; such mixed cases arise, but they are few. Section 31 has examples of a more interesting kind, namely functions F that are continuous but do not come from any density. These are the functions singular in the sense of Lebesgue; the $Q(x)$ describing bold play in gambling (see (7.33)) turns out to be one of them. See Example 31.1.

If X has distribution μ and g is a real function of a real variable,

$$P[g(X) \in A] = P[X \in g^{-1}A] = \mu(g^{-1}A). \tag{20.14}$$

Thus the distribution of $g(X)$ is μg^{-1} in the notation (13.7).

In the case where there is a density, f and F are related by

$$F(x) = \int_{-\infty}^{x} f(t)\,dt. \tag{20.15}$$

Hence f at its continuity points must be the derivative of F. As noted above, if F has a continuous derivative, this derivative can serve as the density f. Suppose that f is continuous and g is increasing, and let $T = g^{-1}$. The distribution function of $g(X)$ is $P[g(X) \le x] = P[X \le T(x)] = F(T(x))$. If T is differentiable, this differentiates to $f(T(x))T'(x)$, which is therefore the density for $g(X)$. If g is decreasing, on the other hand, then $P[g(X) < x] = P[X > T(x)] = 1 - F(T(x))$, and the derivative is equal to $-f(T(x))T'(x) = f(T(x))|T'(x)|$. In either case, $g(X)$ has density

$$\frac{d}{dx} P[g(X) \le x] = f(T(x))|T'(x)|. \tag{20.16}$$

If X has the normal density (20.12) and $a > 0$, (20.16) shows that $aX + b$ has the normal density with parameters $am + b$ and $a\sigma$. Finding the density of $g(X)$ from first principles, as in the argument leading to (20.16), often works even if g is many-to-one:

Example 20.1. If X has the standard normal distribution, then

$$P[X^2 \le x] = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-t^2/2}\,dt = \frac{2}{\sqrt{2\pi}} \int_{0}^{\sqrt{x}} e^{-t^2/2}\,dt$$

for $x > 0$. Hence X^2 has density

$$f(x) = \begin{cases} 0 & \text{if } x \le 0, \\ \dfrac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} & \text{if } x > 0. \end{cases}$$

■

Multidimensional Distributions

For a k-dimensional random vector $X = (X_1, \ldots, X_k)$, the distribution μ (a probability measure on $\mathscr{R}^k$) and the distribution function F (a real function on R^k) are defined by

$$\begin{aligned} \mu(A) &= P[(X_1, \ldots, X_k) \in A], \qquad A \in \mathscr{R}^k, \\ F(x_1, \ldots, x_k) &= P[X_1 \le x_1, \ldots, X_k \le x_k] = \mu(S_x), \end{aligned} \tag{20.17}$$

where $S_x = [y\colon y_i \le x_i,\ i = 1, \ldots, k]$ consists of the points "southwest" of x.

Often μ and F are called the *joint* distribution and *joint* distribution function of $X_1, \ldots, X_k$.

Now F is nondecreasing in each variable, and $\Delta_A F \geq 0$ for bounded rectangles A (see (12.12)). As h decreases to 0, the set

$$S_{x,h} = [y: y_i \leq x_i + h, i = 1, \ldots, k]$$

decreases to S_x, and therefore (Theorem 2.1(ii)) F is continuous from above in the sense that $\lim_{h \downarrow 0} F(x_1 + h, \ldots, x_k + h) = F(x_1, \ldots, x_k)$. Further, $F(x_1, \ldots, x_k) \to 0$ if $x_i \to -\infty$ for some i (the other coordinates held fixed), and $F(x_1, \ldots, x_k) \to 1$ if $x_i \to \infty$ for each i. For any F with these properties there is by Theorem 12.5 a unique probability measure μ on $\mathscr{R}^k$ such that $\mu(A) = \Delta_A F$ for bounded rectangles A, and $\mu(S_x) = F(x)$ for all x.

As h decreases to 0, $S_{x,-h}$ increases to the interior $S_x^\circ = [y: y_i < x_i, i = 1, \ldots, k]$ of S_x, and so

$$\lim_{h \downarrow 0} F(x_1 - h, \ldots, x_k - h) = \mu(S_x^\circ). \tag{20.18}$$

Since F is nondecreasing in each variable, it is continuous at x if and only if it is continuous from below there in the sense that this last limit coincides with $F(x)$. Thus F is continuous at x if and only if $F(x) = \mu(S_x) = \mu(S_x^\circ)$, which holds if and only if the boundary $\partial S_x = S_x - S_x^\circ$ (the y-set where $y_i \leq x_i$ for all i and $y_i = x_i$ for some i) satisfies $\mu(\partial S_x) = 0$. If $k > 1$, F can have discontinuity points even if μ has no point masses: if μ corresponds to a uniform distribution of mass over the segment $B = [(x, 0): 0 < x < 1]$ in the plane ($\mu(A) = \lambda[x: 0 < x < 1, (x, 0) \in A]$), then F is discontinuous at each point of B. This also shows that F can be discontinuous at uncountably many points. On the other hand, for fixed x the boundaries $\partial S_{x,h}$ are disjoint for different values of h, and so (Theorem 10.2(iv)) only countably many of them can have positive μ-measure. Thus x is the limit of points $(x_1 + h, \ldots, x_k + h)$ at which F is continuous: the continuity points of F are dense.

There is always a random vector having a given distribution and distribution function: Take $(\Omega, \mathscr{F}, P) = (R^k, \mathscr{R}^k, \mu)$ and $X(\omega) \equiv \omega$. This is the obvious extension of the construction in the first proof of Theorem 14.1.

The distribution may as for the line be discrete in the sense of having countable support. It may have density f with respect to k-dimensional Lebesgue measure: $\mu(A) = \int_A f(x)\,dx$. As in the case $k = 1$, the distribution μ is more fundamental than the distribution function F, and usually μ is described not by F but by a density or by discrete probabilities.

If X is a k-dimensional random vector and g: $R^k \to R^i$ is measurable, then $g(X)$ is an i-dimensional random vector; if the distribution of X is μ, the distribution of $g(X)$ is μg^{-1}, just as in the case $k = 1$—see (20.14). If g_j: $R^k \to R^1$ is defined by $g_j(x_1, \ldots, x_k) = x_j$, then $g_j(X)$ is X_j, and its distribution $\mu_j = \mu g_j^{-1}$ is given by $\mu_j(A) = \mu[(x_1, \ldots, x_k): x_j \in A] = P[X_j \in A]$ for

$A \in \mathscr{R}^1$. The μ_j are the *marginal distributions* of μ. If μ has a density f in R^k, then μ_j has over the line the density

(20.19)

$$f_j(x) = \int_{R^{k-1}} f(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_k)\, dx_1 \cdots dx_{j-1}\, dx_{j+1} \cdots dx_k,$$

since by Fubini's theorem the right side integrated over A comes to $\mu[(x_1, \ldots, x_k)\colon x_j \in A]$.

Now suppose that g is a one-to-one, continuously differentiable map of V onto U, where U and V are open sets in R^k. Let T be the inverse, and suppose its Jacobian $J(x)$ never vanishes. If X has a density f supported by V, then for $A \subset U$, $P[g(X) \in A] = P[X \in TA] = \int_{TA} f(y)\, dy$, and by (17.10), this equals $\int_A f(Tx)|J(x)|\, dx$. Therefore, $g(X)$ has density

(20.20)
$$d(x) = \begin{cases} f(Tx)|J(x)| & \text{for } x \in U, \\ 0 & \text{for } x \notin U. \end{cases}$$

This is the analogue of (20.16).

Example 20.2. Suppose that (X_1, X_2) has density

$$f(x_1, x_2) = (2\pi)^{-1} \exp\left[-\tfrac{1}{2}\left(x_1^2 + x_2^2\right)\right],$$

and let g be the transformation to polar coordinates. Then U, V, and T are as in Example 17.7. If R and Θ are the polar coordinates of (X_1, X_2), then $(R, \Theta) = g(X_1, X_2)$ has density $(2\pi)^{-1}\rho e^{-\rho^2/2}$ in V. By (20.19), R has density $\rho e^{-\rho^2/2}$ on $(0, \infty)$, and Θ is uniformly distributed over $(0, 2\pi)$. ■

For the normal distribution in R^k, see Section 29.

Independence

Random variables $X_1, \ldots, X_k$ are defined to be independent if the σ-fields $\sigma(X_1), \ldots, \sigma(X_k)$ they generate are independent in the sense of Section 4. This concept for simple random variables was studied extensively in Chapter 1; the general case was touched on in Section 14. Since $\sigma(X_i)$ consists of the sets $[X_i \in H]$ for $H \in \mathscr{R}^1$, $X_1, \ldots, X_k$ are independent if and only if

(20.21) $$P[X_1 \in H_1, \ldots, X_k \in H_k] = P[X_1 \in H_1] \cdots P[X_k \in H_k]$$

for all linear Borel sets $H_1, \ldots, H_k$. The definition (4.10) of independence requires that (20.21) hold also if some of the events $[X_i \in H_i]$ are suppressed on each side, but this only means taking $H_i = R^1$.

Suppose that

$$(20.22) \qquad P[X_1 \le x_1, \ldots, X_k \le x_k] = P[X_1 \le x_1] \cdots P[X_k \le x_k]$$

for all real $x_1, \ldots, x_k$; it then also holds if some of the events $[X_i \le x_i]$ are suppressed on each side (let $x_i \to \infty$). Since the intervals $(-\infty, x]$ form a π-system generating $\mathscr{R}^1$, the sets $[X_i \le x]$ form a π-system generating $\sigma(X_i)$. Therefore, by Theorem 4.2, (20.22) implies that $X_1, \ldots, X_k$ are independent. If, for example, the X_i are integer-valued, it is enough that $P[X_1 = n_1, \ldots, X_k = n_k] = P[X_1 = n_1] \cdots P[X_k = n_k]$ for integral $n_1, \ldots, n_k$ (see (5.9)).

Let $(X_1, \ldots, X_k)$ have distribution μ and distribution function F, and let the X_i have distributions μ_i and distribution functions F_i (the marginals). By (20.21), $X_1, \ldots, X_k$ are independent if and only if μ is product measure in the sense of Section 18:

$$(20.23) \qquad \mu = \mu_1 \times \cdots \times \mu_k.$$

By (20.22), $X_1, \ldots, X_k$ are independent if and only if

$$(20.24) \qquad F(x_1, \ldots, x_k) = F_1(x_1) \cdots F_k(x_k).$$

Suppose that each μ_i has density f_i; by Fubini's theorem, $f_1(y_1) \cdots f_k(y_k)$ integrated over $(-\infty, x_1] \times \cdots \times (-\infty, x_k]$ is just $F_1(x_1) \cdots F_k(x_k)$, so that μ has density

$$(20.25) \qquad f(x) = f_1(x_1) \cdots f_k(x_k)$$

in the case of independence.

If $\mathscr{G}_1, \ldots, \mathscr{G}_k$ are independent σ-fields and X_i is measurable $\mathscr{G}_i$, $i = 1, \ldots, k$, then certainly $X_1, \ldots, X_k$ are independent.

If X_i is a d_i-dimensional random vector, $i = 1, \ldots, k$, then $X_1, \ldots, X_k$ are by definition independent if the σ-fields $\sigma(X_1), \ldots, \sigma(X_k)$ are independent. The theory is just as for random variables: $X_1, \ldots, X_k$ are independent if and only if (20.21) holds for $H_1 \in \mathscr{R}^{d_1}, \ldots, H_k \in \mathscr{R}^{d_k}$. Now $(X_1, \ldots, X_k)$ can be regarded as a random vector of dimension $d = \sum_{i=1}^{k} d_i$; if μ is its distribution in $R^d = R^{d_1} \times \cdots \times R^{d_k}$ and μ_i is the distribution of X_i in R^{d_i}, then, just as before, $X_1, \ldots, X_k$ are independent if and only if $\mu = \mu_1 \times \cdots \times \mu_k$. In none of this need the d_i components of a single X_i be themselves independent random variables.

An infinite collection of random variables or random vectors is by definition independent if each finite subcollection is. The argument following (5.10)

extends from collections of simple random variables to collections of random vectors:

Theorem 20.2. *Suppose that*

$$\begin{array}{lll} X_{11} & X_{12} & \cdots \\ X_{21} & X_{22} & \cdots \\ \vdots & \vdots & \end{array} \tag{20.26}$$

is an independent collection of random vectors. If $\mathscr{F}_i$ is the σ-field generated by the ith row, then $\mathscr{F}_1, \mathscr{F}_2, \ldots$ are independent.

PROOF. Let $\mathscr{A}_i$ consist of the finite intersections of sets of the form $[X_{ij} \in H]$ with H a Borel set in a space of the appropriate dimension, and apply Theorem 4.2. The σ-fields $\mathscr{F}_i = \sigma(\mathscr{A}_i)$, $i = 1, \ldots, n$, are independent for each n, and the result follows. ■

Each row of (20.26) may be finite or infinite, and there may be finitely or infinitely many rows. As a matter of fact, rows may be uncountable and there may be uncountably many of them.

Suppose that X and Y are independent random vectors with distributions μ and ν in R^j and R^k. Then (X, Y) has distribution $\mu \times \nu$ in $R^j \times R^k = R^{j+k}$. Let x range over R^j and y over R^k. By Fubini's theorem,

$$(\mu \times \nu)(B) = \int_{R^j} \nu[y: (x, y) \in B]\mu(dx), \qquad B \in \mathscr{R}^{j+k}. \tag{20.27}$$

Replace B by $(A \times R^k) \cap B$, where $A \in \mathscr{R}^j$ and $B \in \mathscr{R}^{j+k}$. Then (20.27) reduces to

$$(\mu \times \nu)\big((A \times R^k) \cap B\big) = \int_A \nu[y: (x, y) \in B]\mu(dx), \tag{20.28}$$

$$A \in \mathscr{R}^j, \quad B \in \mathscr{R}^{j+k}.$$

If $B_x = [y: (x, y) \in B]$ is the x-section of B, so that $B_x \in \mathscr{R}^k$ (Theorem 18.1), then $P[(x, Y) \in B] = P[\omega: (x, Y(\omega)) \in B] = P[\omega: Y(\omega) \in B_x] = \nu(B_x)$. Expressing the formulas in terms of the random vectors themselves gives this result:

Theorem 20.3. *If X and Y are independent random vectors with distributions μ and ν in R^j and R^k, then*

$$P[(X, Y) \in B] = \int_{R^j} P[(x, Y) \in B]\mu(dx), \qquad B \in \mathscr{R}^{j+k}, \tag{20.29}$$

and

$$P[X\in A,(X,Y)\in B]=\int_A P[(x,Y)\in B]\mu(dx), \tag{20.30}$$

$$A\in\mathscr{R}^j,\quad B\in\mathscr{R}^{j+k}.$$

Example 20.3. Suppose that X and Y are independent exponentially distributed random variables. By (20.29), $P[Y/X\ge z]=\int_0^\infty P[Y/x\ge z]\alpha e^{-\alpha x}\,dx=\int_0^\infty e^{-\alpha xz}\alpha e^{-\alpha x}\,dx=(1+z)^{-1}$. Thus Y/X has density $(1+z)^{-2}$ for $z\ge 0$. Since $P[X\ge z_1,\ Y/X\ge z_2]=\int_{z_1}^\infty P[Y/x\ge z_2]\alpha e^{-\alpha x}\,dx$ by (20.30), the joint distribution of X and Y/X can be calculated as well. ■

The formulas (20.29) and (20.30) are constantly applied as in this example. There is no virtue in making an issue of each case, however, and the appeal to Theorem 20.3 is usually silent.

Example 20.4. Here is a more complicated argument of the same sort. Let $X_1,\ldots,X_n$ be independent random variables, each uniformly distributed over $[0,t]$. Let Y_k be the kth smallest among the X_i, so that $0\le Y_1\le\cdots\le Y_n\le t$. The X_i divide $[0,t]$ into $n+1$ subintervals of lengths $Y_1, Y_2-Y_1,\ldots,Y_n-Y_{n-1}, t-Y_n$; let M be the maximum of these lengths. Define $\psi_n(t,a)=P[M\le a]$. The problem is to show that

$$\psi_n(t,a)=\sum_{k=0}^{n+1}(-1)^k\binom{n+1}{k}\left(1-k\frac{a}{t}\right)_+^n, \tag{20.31}$$

where $x_+=(x+|x|)/2$ denotes positive part.

Separate consideration of the possibilities $0\le a\le t/2$, $t/2\le a\le t$, and $t\le a$ disposes of the case $n=1$. Suppose it is shown that the probability $\psi_n(t,a)$ satisfies the recursion

$$\psi_n(t,a)=n\int_0^a\psi_{n-1}(t-x,a)\left(\frac{t-x}{t}\right)^{n-1}\frac{dx}{t}. \tag{20.32}$$

Now (as follows by an integration together with Pascal's identity for binomial coefficients) the right side of (20.31) satisfies this same recursion, and so it will follow by induction that (20.31) holds for all n.

In intuitive form, the argument for (20.32) is this: If $[M\le a]$ is to hold, the smallest of the X_i must have some value x in $[0,a]$. If X_1 is the smallest of the X_i, then $X_2,\ldots,X_n$ must all lie in $[x,t]$ and divide it into subintervals of length at most a; the probability of this is $(1-x/t)^{n-1}\psi_{n-1}(t-x,a)$, because $X_2,\ldots,X_n$ have probability $(1-x/t)^{n-1}$ of all lying in $[x,t]$, and if they do, they are independent and uniformly distributed there. Now (20.32) results from integrating with respect to the density for X_1 and multiplying by n to allow for the fact that any of $X_1,\ldots,X_n$ may be the smallest.

To make this argument rigorous, apply (20.30) for $j=1$ and $k=n-1$. Let A be the interval $[0,a]$, and let B consist of the points $(x_1,\ldots,x_n)$ for which $0\le x_i\le t$, x_1 is the minimum of $x_1,\ldots,x_n$, and $x_2,\ldots,x_n$ divide $[x_1,t]$ into subintervals of length at most a. Then $P[X_1=\min X_i,\ M\le a]=P[X_1\in A,(X_1,\ldots,X_n)\in B]$. Take X_1 for

X and $(X_2, \ldots, X_n)$ for Y in (20.30). Since X_1 has density $1/t$,

$$P[X_1 = \min X_i,\ M \le a] = \int_0^a P[(x, X_2, \ldots, X_n) \in B]\frac{dx}{t}. \tag{20.33}$$

If C is the event that $x \le X_i \le t$ for $2 \le i \le n$, then $P(C) = (1 - x/t)^{n-1}$. A simple calculation shows that $P[X_i - x \le s_i,\ 2 \le i \le n | C] = \prod_{i=2}^n (s_i/(t-x))$; in other words, given C, the random variables $X_2 - x, \ldots, X_n - x$ are conditionally independent and uniformly distributed over $[0, t-x]$. Now $X_2, \ldots, X_n$ are random variables on some probability space $(\Omega, \mathscr{F}, P)$; replacing P by $P(\cdot|C)$ shows that the integrand in (20.33) is the same as that in (20.32). The same argument holds with the index 1 replaced by any k $(1 \le k \le n)$, which gives (20.32). (The events $[X_k = \min X_i,\ Y \le a]$ are not disjoint, but any two intersect in a set of probability 0.) ■

Sequences of Random Variables

Theorem 5.3 extends to general distributions μ_n.

Theorem 20.4. *If $\{\mu_n\}$ is a finite or infinite sequence of probability measures on $\mathscr{R}^1$, there exists on some probability space $(\Omega, \mathscr{F}, P)$ an independent sequence $\{X_n\}$ of random variables such that X_n has distribution μ_n.*

PROOF. By Theorem 5.3 there exists on some probability space an independent sequence $Z_1, Z_2, \ldots$ of random variables assuming the values 0 and 1 with probabilities $P[Z_n = 0] = P[Z_n = 1] = \frac{1}{2}$. As a matter of fact, Theorem 5.3 is not needed: take the space to be the unit interval and the $Z_n(\omega)$ to be the digits of the dyadic expansion of ω—the functions $d_n(\omega)$ of Sections and 1 and 4.

Relabel the countably many random variables Z_n so that they form a double array,

$$\begin{matrix} Z_{11} & Z_{12} & \cdots \\ Z_{21} & Z_{22} & \cdots \\ \vdots & \vdots & \end{matrix}$$

All the Z_{nk} are independent. Put $U_n = \sum_{k=1}^\infty Z_{nk} 2^{-k}$. The series certainly converges, and U_n is a random variable by Theorem 13.4. Further, $U_1, U_2, \ldots$ is, by Theorem 20.2, an independent sequence.

Now $P[Z_{ni} = z_i,\ 1 \le i \le k] = 2^{-k}$ for each sequence $z_1, \ldots, z_k$ of 0's and 1's; hence the 2^k possible values $j2^{-k}$, $0 \le j < 2^k$, of $S_{nk} = \sum_{i=1}^k Z_{ni} 2^{-i}$ all have probability 2^{-k}. If $0 \le x < 1$, the number of the $j2^{-k}$ that lie in $[0, x]$ is $\lfloor 2^k x \rfloor + 1$, and therefore $P[S_{nk} \le x] = (\lfloor 2^k x \rfloor + 1)/2^k$. Since $S_{nk}(\omega) \uparrow U_n(\omega)$ as $k \uparrow \infty$, it follows that $[S_{nk} \le x] \downarrow [U_n \le x]$ as $k \uparrow \infty$, and so $P[U_n \le x] = \lim_k P[S_{nk} \le x] = \lim_k (\lfloor 2^k x \rfloor + 1)/2^k = x$ for $0 \le x < 1$. Thus U_n is uniformly distributed over the unit interval.

The construction thus far establishes the existence of an independent sequence of random variables U_n each uniformly distributed over $[0, 1]$. Let F_n be the distribution function corresponding to μ_n, and put $\varphi_n(u) = \inf[x: u \le F_n(x)]$ for $0 < u < 1$. This is the inverse used in Section 14—see (14.5). Set $\varphi_n(u) = 0$, say, for u outside $(0, 1)$, and put $X_n(\omega) = \varphi_n(U_n(\omega))$. Since $\varphi_n(u) \le x$ if and only if $u \le F_n(x)$—see the argument following (14.5)—$P[X_n \le x] = P[U_n \le F_n(x)] = F_n(x)$. Thus X_n has distribution function F_n. And by Theorem 20.2, $X_1, X_2, \ldots$ are independent. ■

This theorem of course includes Theorem 5.3 as a special case, and its proof does not depend on the earlier result. Theorem 20.4 is a special case of Kolmogorov's existence theorem in Section 36.

Convolution

Let X and Y be independent random variables with distributions μ and ν. Apply (20.27) and (20.29) to the planar set $B = [(x, y): x + y \in H]$ with $H \in \mathscr{R}^1$:

$$P[X + Y \in H] = \int_{-\infty}^{\infty} \nu(H - x)\mu(dx) \tag{20.34}$$

$$= \int_{-\infty}^{\infty} P[Y \in H - x]\mu(dx).$$

The *convolution* of μ and ν is the measure $\mu * \nu$ defined by

$$(\mu * \nu)(H) = \int_{-\infty}^{\infty} \nu(H - x)\mu(dx), \qquad H \in \mathscr{R}^1. \tag{20.35}$$

If X and Y are independent and have distributions μ and ν, (20.34) shows that $X + Y$ has distribution $\mu * \nu$. Since addition of random variables is commutative and associative, the same is true of convolution: $\mu * \nu = \nu * \mu$ and $\mu * (\nu * \eta) = (\mu * \nu) * \eta$.

If F and G are the distribution functions corresponding to μ and ν, the distribution function corresponding to $\mu * \nu$ is denoted $F * G$. Taking $H = (-\infty, y]$ in (20.35) shows that

$$(F * G)(y) = \int_{-\infty}^{\infty} G(y - x)\, dF(x). \tag{20.36}$$

(See (17.22) for the notation $dF(x)$.) If G has density g, then $G(y - x) = \int_{-\infty}^{y-x} g(s)\, ds = \int_{-\infty}^{y} g(t - x)\, dt$, and so the right side of (20.36) is $\int_{-\infty}^{y} [\int_{-\infty}^{\infty} g(t - x)\, dF(x)]\, dt$ by Fubini's theorem. Thus $F * G$ has density $F * g$, where

$$(F * g)(y) = \int_{-\infty}^{\infty} g(y - x)\, dF(x); \tag{20.37}$$

this holds if G has density g. If, in addition, F has density f, (20.37) is denoted $f * g$ and reduces by (16.12) to

$$(20.38) \qquad (f * g)(y) = \int_{-\infty}^{\infty} g(y-x)f(x)\,dx.$$

This defines convolution for densities, and $\mu * \nu$ has density $f * g$ if μ and ν have densities f and g. The formula (20.38) can be used for many explicit calculations.

Example 20.5. Let $X_1, \ldots, X_k$ be independent random variables, each with the exponential density (20.10). Define g_k by

$$(20.39) \qquad g_k(x) = \alpha \frac{(\alpha x)^{k-1}}{(k-1)!} e^{-\alpha x}, \qquad x \geq 0, \quad k = 1, 2, \ldots;$$

put $g_k(x) = 0$ for $x \leq 0$. Now

$$(g_{k-1} * g_1)(y) = \int_0^y g_{k-1}(y-x)g_1(x)\,dx,$$

which reduces to $g_k(y)$. Thus $g_k = g_{k-1} * g_1$, and since g_1 coincides with (20.10), it follows by induction that the sum $X_1 + \cdots + X_k$ has density g_k. The corresponding distribution function is

$$(20.40) \quad G_k(x) = 1 - e^{-\alpha x} \sum_{i=0}^{k-1} \frac{(\alpha x)^i}{i!} = \sum_{i=k}^{\infty} e^{-\alpha x} \frac{(\alpha x)^i}{i!}, \qquad x \geq 0,$$

as follows by differentiation. ∎

Example 20.6. Suppose that X has the normal density (20.12) with $m = 0$ and that Y has the same density with τ in place of σ. If X and Y are independent, then $X + Y$ has density

$$\frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} \exp\left[-\frac{(y-x)^2}{2\sigma^2} - \frac{x^2}{2\tau^2} \right] dx.$$

A change of variable $u = x(\sigma^2 + \tau^2)^{1/2}/\sigma\tau$ reduces this to

$$\frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(u - y\frac{\tau/\sigma}{\sqrt{\sigma^2+\tau^2}} \right)^2 - \frac{y^2}{2(\sigma^2+\tau^2)} \right] du$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} e^{-y^2/2(\sigma^2+\tau^2)}.$$

Thus $X+Y$ has the normal density with $m=0$ and with $\sigma^2+\tau^2$ in place of σ^2. ■

If μ and ν are arbitrary finite measures on the line, their convolution is defined by (20.35) even if they are not probability measures.

Convergence in Probability

Random variables X_n converge in probability to X, written $X_n \to_P X$, if

$$\lim_n P[|X_n - X| \geq \epsilon] = 0 \tag{20.41}$$

for each positive ϵ.[†] This is the same as (5.7), and the proof of Theorem 5.2 carries over without change (see also Example 5.4).

Theorem 20.5. (i) *If $X_n \to X$ with probability 1, then $X_n \to_P X$.*

(ii) *A necessary and sufficient condition for $X_n \to_P X$ is that each subsequence $\{X_{n_k}\}$ contain a further subsequence $\{X_{n_{k(i)}}\}$ such that $X_{n_{k(i)}} \to X$ with probability 1 as $i \to \infty$.*

PROOF. Only part (ii) needs proof. If $X_n \to_P X$, then given $\{n_k\}$, choose a subsequence $\{n_{k(i)}\}$ so that $k \geq k(i)$ implies that $P[|X_{n_k} - X| \geq i^{-1}] < 2^{-i}$. By the first Borel–Cantelli lemma there is probability 1 that $|X_{n_{k(i)}} - X| < i^{-1}$ for all but finitely many i. Therefore, $\lim_i X_{n_{k(i)}} = X$ with probability 1.

If X_n does not converge to X in probability, there is some positive ϵ for which $P[|X_{n_k} - X| \geq \epsilon] > \epsilon$ holds along some sequence $\{n_k\}$. No subsequence of $\{X_{n_k}\}$ can converge to X in probability, and hence none can converge to X with probability 1. ■

It follows from (ii) that if $X_n \to_P X$ and $X_n \to_P Y$, then $X = Y$ with probability 1. It follows further that if f is continuous and $X_n \to_P X$, then $f(X_n) \to_P f(X)$.

In nonprobabilistic contexts, convergence in probability becomes *convergence in measure*: If f_n and f are real measurable functions on a measure space $(\Omega, \mathscr{F}, \mu)$, and if $\mu[\omega: |f(\omega) - f_n(\omega)| \geq \epsilon] \to 0$ for each $\epsilon > 0$, then f_n converges in measure to f.

The Glivenko–Cantelli Theorem*

The *empirical distribution function* for random variables $X_1, \ldots, X_n$ is the distribution function $F_n(x, \omega)$ with a jump of n^{-1} at each $X_k(\omega)$:

$$F_n(x, \omega) = \frac{1}{n} \sum_{k=1}^{n} I_{(-\infty, x]}(X_k(\omega)). \tag{20.42}$$

[†]This is often expressed $\operatorname{p\,lim}_n X_n = X$.

*This topic may be omitted.

If the X_k have a common unknown distribution function $F(x)$, then $F_n(x, \omega)$ is its natural estimate. The estimate has the right limiting behavior, according to the *Glivenko–Cantelli theorem*:

Theorem 20.6. *Suppose that $X_1, X_2, \ldots$ are independent and have a common distribution function F; put $D_n(\omega) = \sup_x |F_n(x, \omega) - F(x)|$. Then $D_n \to 0$ with probability* 1.

For each x, $F_n(x, \omega)$ as a function of ω is a random variable. By right continuity, the supremum above is unchanged if x is restricted to the rationals, and therefore D_n is a random variable.

The summands in (20.42) are independent, identically distributed simple random variables, and so by the strong law of large numbers (Theorem 6.1), for each x there is a set A_x of probability 0 such that

$$\lim_n F_n(x, \omega) = F(x) \tag{20.43}$$

for $\omega \notin A_x$. But Theorem 20.6 says more, namely that (20.43) holds for ω outside some set A of probability 0, where A does not depend on x—as there are uncountably many of the sets A_x, it is conceivable a priori that their union might necessarily have positive measure. Further, the convergence in (20.43) is uniform in x. Of course, the theorem implies that with probability 1 there is weak convergence $F_n(x, \omega) \Rightarrow F(x)$ in the sense of Section 14.

PROOF OF THE THEOREM. As already observed, the set A_x where (20.43) fails has probability 0. Another application of the strong law of large numbers, with $I_{(-\infty, x)}$ in place of $I_{(-\infty, x]}$ in (20.42), shows that (see (20.5)) $\lim_n F_n(x-, \omega) = F(x-)$ except on a set B_x of probability 0. Let $\varphi(u) = \inf[x: u \le F(x)]$ for $0 < u < 1$ (see (14.5)), and put $x_{m,k} = \varphi(k/m)$, $m \ge 1$, $1 \le k \le m$. It is not hard to see that $F(\varphi(u)-) \le u \le F(\varphi(u))$; hence $F(x_{m,k}-) - F(x_{m,k-1}) \le m^{-1}$, $F(x_{m,1}-) \le m^{-1}$, and $F(x_{m,m}) \ge 1 - m^{-1}$. Let $D_{m,n}(\omega)$ be the maximum of the quantities $|F_n(x_{m,k}, \omega) - F(x_{m,k})|$ and $|F_n(x_{m,k}-, \omega) - F(x_{m,k}-)|$ for $k = 1, \ldots, m$.

If $x_{m,k-1} \le x < x_{m,k}$, then $F_n(x, \omega) \le F_n(x_{m,k}-, \omega) \le F(x_{m,k}-) + D_{m,n}(\omega) \le F(x) + m^{-1} + D_{m,n}(\omega)$ and $F_n(x, \omega) \ge F_n(x_{m,k-1}, \omega) \ge F(x_{m,k-1}) - D_{m,n}(\omega) \ge F(x) - m^{-1} - D_{m,n}(\omega)$. Together with similar arguments for the cases $x < x_{m,1}$ and $x \ge x_{m,m}$, this shows that

$$D_n(\omega) \le D_{m,n}(\omega) + m^{-1}. \tag{20.44}$$

If ω lies outside the union A of all the $A_{x_{mk}}$ and $B_{x_{mk}}$, then $\lim_n D_{m,n}(\omega) = 0$ and hence $\lim_n D_n(\omega) = 0$ by (20.44). But A has probability 0. ■

PROBLEMS

20.1. 2.11↑ A necessary and sufficient condition for a σ-field $\mathscr{G}$ to be countably generated is that $\mathscr{G}=\sigma(X)$ for some random variable X. *Hint:* If $\mathscr{G}=\sigma(A_1, A_2, \ldots)$, consider $X=\sum_{k=1}^{\infty} f(I_{A_k})/10^k$, where $f(x)$ is 4 for $x=0$ and 5 for $x\neq 0$.

20.2. If X is a positive random variable with density f, then X^{-1} has density $f(1/x)/x^2$. Prove this by (20.16) and by a direct argument.

20.3. Suppose that a two-dimensional distribution function F has a continuous density f. Show that $f(x, y)=\partial^2 F(x, y)/\partial x\,\partial y$.

20.4. The construction in Theorem 20.4 requires only Lebesgue measure on the unit interval. Use the theorem to prove the existence of Lebesgue measure on R^k. First construct λ_k restricted to $(-n, n]\times\cdots\times(-n, n]$, and then pass to the limit $(n\to\infty)$. The idea is to argue from first principles, and not to use previous constructions, such as those in Theorems 12.5 and 18.2.

20.5. Suppose that A, B, and C are positive, independent random variables with distribution function F. Show that the quadratic Az^2+Bz+C has real zeros with probability $\int_0^\infty\int_0^\infty F(x^2/4y)\,dF(x)\,dF(y)$.

20.6. Show that $X_1, X_2, \ldots$ are independent if $\sigma(X_1, \ldots, X_{n-1})$ and $\sigma(X_n)$ are independent for each n.

20.7. Let $X_0, X_1, \ldots$ be a persistent, irreducible Markov chain, and for a fixed state j let $T_1, T_2, \ldots$ be the times of the successive passages through j. Let $Z_1=T_1$ and $Z_n=T_n-T_{n-1}$, $n\geq 2$. Show that $Z_1, Z_2, \ldots$ are independent and that $P[Z_n=k]=f_{jj}^{(k)}$ for $n\geq 2$.

20.8. *Ranks and records.* Let $X_1, X_2, \ldots$ be independent random variables with a common continuous distribution function. Let B be the ω-set where $X_m(\omega)=X_n(\omega)$ for some pair m, n of distinct integers, and show that $P(B)=0$. Remove B from the space Ω on which the X_n are defined. This leaves the joint distributions of the X_n unchanged and makes ties impossible.

Let $T^{(n)}(\omega)=(T_1^{(n)}(\omega), \ldots, T_n^{(n)}(\omega))$ be that permutation $(t_1, \ldots, t_n)$ of $(1, \ldots, n)$ for which $X_{t_1}(\omega)<X_{t_2}(\omega)<\cdots<X_{t_n}(\omega)$. Let Y_n be the rank of X_n among $X_1, \ldots, X_n$: $Y_n=r$ if and only if $X_i<X_n$ for exactly $r-1$ values of i preceding n.

(a) Show that $T^{(n)}$ is uniformly distributed over the $n!$ permutations.

(b) Show that $P[Y_n=r]=1/n$, $1\leq r\leq n$.

(c) Show that Y_k is measurable $\sigma(T^{(n)})$ for $k\leq n$.

(d) Show that $Y_1, Y_2, \ldots$ are independent.

20.9. ↑ *Record values.* Let A_n be the event that a *record* occurs at time n: $\max_{k<n} X_k<X_n$.

(a) Show that $A_1, A_2, \ldots$ are independent and $P(A_n)=1/n$.

(b) Show that no record stands forever.

(c) Let N_n be the time of the first record after time n. Show that $P[N_n=n+k]=n(n+k-1)^{-1}(n+k)^{-1}$.

20.10. Use Fubini's theorem to prove that convolution of finite measures is commutative and associative.

20.11. Suppose that X and Y are independent and have densities. Use (20.20) to find the joint density for $(X+Y, X)$ and then use (20.19) to find the density for $X+Y$. Check with (20.38).

20.12. If $F(x-\epsilon) < F(x+\epsilon)$ for all positive ϵ, then x is a *point of increase* of F (see Problem 12.9). If $F(x-) < F(x)$, then x is an *atom* of F.

(a) Show that, if x and y are points of increase of F and G, then $x+y$ is a point of increase of $F * G$.

(b) Show that, if x and y are atoms of F and G, then $x+y$ is an atom of $F * G$.

20.13. Suppose that μ and ν consist of masses α_n and β_n at n, $n = 0, 1, 2, \ldots$. Show that $\mu * \nu$ consists of a mass of $\sum_{k=0}^{n} \alpha_k \beta_{n-k}$ at n, $n = 0, 1, 2, \ldots$. Show that two Poisson distributions (the parameters may differ) convolve to a Poisson distribution.

20.14. The *Cauchy* distribution has density

$$c_u(x) = \frac{1}{\pi} \frac{u}{u^2 + x^2}, \qquad -\infty < x < \infty, \tag{20.45}$$

for $u > 0$. (By (17.9), the density integrates to 1.)

(a) Show that $c_u * c_v = c_{u+v}$. *Hint:* Expand the convolution integrand in partial fractions.

(b) Show that, if $X_1, \ldots, X_n$ are independent and have density c_u, then $(X_1 + \cdots + X_n)/n$ has density c_u as well.

20.15. ↑ (a) Show that, if X and Y are independent and have the standard normal density, then X/Y has the Cauchy density with $u = 1$.

(b) Show that, if X has the uniform distribution over $(-\pi/2, \pi/2)$, then $\tan X$ has the Cauchy distribution with $u = 1$.

20.16. 18.18↑ Let $X_1, \ldots, X_n$ be independent, each having the standard normal distribution. Show that

$$\chi_n^2 = X_1^2 + \cdots + X_n^2$$

has density

$$\frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} \tag{20.46}$$

over $(0, \infty)$. This is called the *chi-squared distribution with n degrees of freedom*.

20.17. ↑ The *gamma distribution* has density

$$f(x;\alpha,u)=\frac{\alpha^u}{\Gamma(u)}x^{u-1}e^{-\alpha x} \tag{20.47}$$

over $(0,\infty)$ for positive parameters α and u. Check that (20.47) integrates to 1. Show that

$$f(\cdot;\alpha,u)*f(\cdot;\alpha,v)=f(\cdot;\alpha,u+v). \tag{20.48}$$

Note that (20.46) is $f(x;\frac{1}{2},n/2)$, and from (20.48) deduce again that (20.46) is the density of χ_n^2. Note that the exponential density (20.10) is $f(x;\alpha,1)$, and from (20.48) deduce (20.39) once again.

20.18. ↑ Let $N, X_1, X_2, \ldots$ be independent, where $P[N=n]=q^{n-1}p$, $n\geq 1$, and each X_k has the exponential density $f(x;\alpha,1)$. Show that $X_1+\cdots+X_N$ has density $f(x;\alpha p,1)$.

20.19. Let $A_{nm}(\epsilon)=[|Z_k-Z|<\epsilon,\ n\leq k\leq m]$. Show that $Z_n\to Z$ with probability 1 if and only if $\lim_n \lim_m P(A_{nm}(\epsilon))=1$ for all positive ϵ, whereas $Z_n\to_P Z$ if and only if $\lim_n P(A_{nn}(\epsilon))=1$ for all positive ϵ.

20.20. (a) Suppose that $f\colon R^2\to R^1$ is continuous. Show that $X_n\to_P X$ and $Y_n\to_P Y$ imply $f(X_n,Y_n)\to_P f(X,Y)$.

(b) Show that addition and multiplication preserve convergence in probability.

20.21. Suppose that the sequence $\{X_n\}$ is *fundamental in probability* in the sense that for ϵ positive there exists an N_ϵ such that $P[|X_m-X_n|>\epsilon]<\epsilon$ for $m,n>N_\epsilon$.

(a) Prove there is a subsequence $\{X_{n_k}\}$ and a random variable X such that $\lim_k X_{n_k}=X$ with probability 1. *Hint:* Choose increasing n_k such that $P[|X_m-X_n|>2^{-k}]<2^{-k}$ for $m,n\geq n_k$. Analyze $P[|X_{n_{k+1}}-X_{n_k}|>2^{-k}]$.

(b) Show that $X_n\to_P X$.

20.22. (a) Suppose that $X_1\leq X_2\leq\cdots$ and that $X_n\to_P X$. Show that $X_n\to X$ with probability 1.

(b) Show by example that in an infinite measure space functions can converge almost everywhere without converging in measure.

20.23. If $X_n\to 0$ with probability 1, then $n^{-1}\sum_{k=1}^n X_k\to 0$ with probability 1 by the standard theorem on Cesàro means [A30]. Show by example that this is not so if convergence with probability 1 is replaced by convergence in probability.

20.24. 2.19↑ (a) Show that in a discrete probability space convergence in probability is equivalent to convergence with probability 1.

(b) Show that discrete spaces are essentially the only ones where this equivalence holds: Suppose that P has a nonatomic part in the sense that there is a set A such that $P(A)>0$ and $P(\cdot|A)$ is nonatomic. Construct random variables X_n such that $X_n\to_P 0$ but X_n does not converge to 0 with probability 1.

20.25. 20.21 20.24 ↑ Let $d(X,Y)$ be the infimum of those positive ϵ for which $P[|X-Y|\geq\epsilon]\leq\epsilon$.

(a) Show that $d(X,Y)=0$ if and only if $X=Y$ with probability 1. Identify random variables that are equal with probability 1, and show that d is a metric on the resulting space.

(b) Show that $X_n\to_P X$ if and only if $d(X_n,X)\to 0$.

(c) Show that the space is complete.

(d) Show that in general there is no metric d_0 on this space such that $X_n\to X$ with probability 1 if and only if $d_0(X_n,X)\to 0$.

20.26. Construct in R^k a random variable X that is uniformly distributed over the surface of the unit sphere in the sense that $|X|=1$ and UX has the same distribution as X for orthogonal transformations U. *Hint:* Let Z be uniformly distributed in the unit ball in R^k, define $\psi(x)=x/|x|$ ($\psi(0)=(1,0,\ldots,0)$, say), and take $X=\psi(Z)$.

20.27. ↑ Let Θ and Φ be the longitude and latitude of a random point on the surface of the unit sphere in R^3. Show that Θ and Φ are independent, Θ is uniformly distributed over $[0,2\pi)$, and Φ is distributed over $[-\pi/2,+\pi/2]$ with density $\frac{1}{2}\cos\phi$.

SECTION 21. EXPECTED VALUES

Expected Value as Integral

The expected value of a random variable X on $(\Omega,\mathcal{F},P)$ is the integral of X with respect to the measure P:

$$E[X]=\int X\,dP=\int_\Omega X(\omega)\,P(d\omega).$$

All the definitions, conventions, and theorems of Chapter 3 apply. For nonnegative X, $E[X]$ is always defined (it may be infinite); for the general X, $E[X]$ is defined, or X has an expected value, if at least one of $E[X^+]$ and $E[X^-]$ is finite, in which case $E[X]=E[X^+]-E[X^-]$; and X is integrable if and only if $E[|X|]<\infty$. The integral $\int_A X\,dP$ over a set A is defined, as before, as $E[I_A X]$. In the case of simple random variables, the definition reduces to that used in Sections 5 through 9.

Expected Values and Limits

The theorems on integration to the limit in Section 16 apply. A useful fact: If random variables X_n are dominated by an integrable random variable, or if they are uniformly integrable, then $E[X_n]\to E[X]$ follows if X_n converges to X *in probability*—convergence with probability 1 is not necessary. This follows easily from Theorem 20.5.

Expected Values and Distributions

Suppose that X has distribution μ. If g is a real function of a real variable, then by the change-of-variable formula (16.17),

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\mu(dx). \tag{21.1}$$

(In applying (16.17), replace $T\colon \Omega \to \Omega'$ by $X\colon \Omega \to R^1$, μ by P, μT^{-1} by μ, and f by g.) This formula holds in the sense explained in Theorem 16.13: It holds in the nonnegative case, so that

$$E[|g(X)|] = \int_{-\infty}^{\infty} |g(x)|\mu(dx); \tag{21.2}$$

if one side is infinite, then so is the other. And if the two sides of (21.2) are finite, then (21.1) holds.

If μ is discrete and $\mu\{x_1, x_2, \ldots\} = 1$, then (21.1) becomes (use Theorem 16.9)

$$E[g(X)] = \sum_r g(x_r)\mu\{x_r\}. \tag{21.3}$$

If X has density f, then (21.1) becomes (use Theorem 16.11)

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx. \tag{21.4}$$

If F is the distribution function of X and μ, (21.1) can be written $E[g(X)] = \int_{-\infty}^{\infty} g(x)\,dF(x)$ in the notation (17.22).

Moments

By (21.2), μ and F determine all the *absolute moments* of X:

$$E\left[|X|^k\right] = \int_{-\infty}^{\infty} |x|^k \mu(dx) = \int_{-\infty}^{\infty} |x|^k\,dF(x), \qquad k = 1, 2, \ldots. \tag{21.5}$$

Since $j \le k$ implies that $|x|^j \le 1 + |x|^k$, if X has a finite absolute moment of order k, then it has finite absolute moments of orders $1, 2, \ldots, k-1$ as well. For each k for which (2.15) is finite, X has kth *moment*

$$E\left[X^k\right] = \int_{-\infty}^{\infty} x^k \mu(dx) = \int_{-\infty}^{\infty} x^k\,dF(x). \tag{21.6}$$

These quantities are also referred to as the moments of μ and of F. They can be computed by (21.3) and (21.4) in the appropriate circumstances.

Example 21.1. Consider the normal density (20.12) with $m = 0$ and $\sigma = 1$. For each k, $x^k e^{-x^2/2}$ goes to 0 exponentially as $x \to \pm\infty$, and so finite moments of all orders exist. Integration by parts shows that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{-x^2/2}\, dx = \frac{k-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{k-2} e^{-x^2/2}\, dx, \qquad k = 2, 3, \ldots.$$

(Apply (18.16) to $g(x) = x^{k-2}$ and $f(x) = xe^{-x^2/2}$, and let $a \to -\infty$, $b \to \infty$.) Of course, $E[X] = 0$ by symmetry and $E[X^0] = 1$. It follows by induction that

$$E\left[X^{2k}\right] = 1 \times 3 \times 5 \times \cdots \times (2k-1), \qquad k = 1, 2, \ldots, \tag{21.7}$$

and that the odd moments all vanish. ■

If the first two moments of X are finite and $E[X] = m$, then just as in Section 5, the *variance* is

$$\begin{aligned} \operatorname{Var}[X] &= E\left[(X-m)^2\right] = \int_{-\infty}^{\infty} (x-m)^2 \mu(dx) \\ &= E\left[X^2\right] - m^2. \end{aligned} \tag{21.8}$$

From Example 21.1 and a change of variable, it follows that a random variable with the normal density (20.12) has mean m and variance σ^2.

Consider for nonnegative X the relation

$$E[X] = \int_0^{\infty} P[X > t]\, dt = \int_0^{\infty} P[X \geq t]\, dt. \tag{21.9}$$

Since $P[X = t]$ can be positive for at most countably many values of t, the two integrands differ only on a set of Lebesgue measure 0 and hence the integrals are the same. For X simple and nonnegative, (21.9) was proved in Section 5; see (5.29). For the general nonnegative X, let X_n be simple random variables for which $0 \leq X_n \uparrow X$ (see (20.1)). By the monotone convergence theorem $E[X_n] \uparrow E[X]$; moreover, $P[X_n > t] \uparrow P[X > t]$, and therefore $\int_0^{\infty} P[X_n > t]\, dt \uparrow \int_0^{\infty} P[X > t]\, dt$, again by the monotone convergence theorem. Since (21.9) holds for each X_n, a passage to the limit establishes (21.9) for X itself. Note that both sides of (21.9) may be infinite. If the integral on the right is finite, then X is integrable.

Replacing X by $XI_{[X>\alpha]}$ leads from (21.9) to

$$\int_{[X>\alpha]} X\, dP = \alpha P[X > \alpha] + \int_{\alpha}^{\infty} P[X > t]\, dt, \qquad \alpha \geq 0. \tag{21.10}$$

As long as $\alpha \geq 0$, this holds even if X is not nonnegative.

Inequalities

Since the final term in (21.10) is nonnegative, $\alpha P[X \geq \alpha] \leq \int_{[X \geq \alpha]} X\,dP \leq E[X]$. Thus

$$P[X \geq \alpha] \leq \frac{1}{\alpha} \int_{[X \geq \alpha]} X\,dP \leq \frac{1}{\alpha} E[X], \qquad \alpha > 0, \tag{21.11}$$

for nonnegative X. As in Section 5, there follow the inequalities

$$P[|X| \geq \alpha] \leq \frac{1}{\alpha^k} \int_{[|X| \geq \alpha]} |X|^k\,dP \leq \frac{1}{\alpha^k} E[|X|^k]. \tag{21.12}$$

It is the inequality between the two extreme terms here that usually goes under the name of Markov; but the left-hand inequality is often useful, too. As a special case there is Chebyshev's inequality,

$$P[|X - m| \geq \alpha] \leq \frac{1}{\alpha^2} \operatorname{Var}[X] \tag{21.13}$$

($m = E[X]$).

Jensen's inequality

$$\varphi(E[X]) \leq E[\varphi(X)] \tag{21.14}$$

holds if φ is convex on an interval containing the range of X and if X and $\varphi(X)$ both have expected values. To prove it, let $l(x) = ax + b$ be a supporting line through $(E[X], \varphi(E[X]))$—a line lying entirely under the graph of φ [A33]. Then $aX(\omega) + b \leq \varphi(X(\omega))$, so that $aE[X] + b \leq E[\varphi(X)]$. But the left side of this inequality is $\varphi(E[X])$.

Hölder's inequality is

$$E[|XY|] \leq E^{1/p}[|X|^p] E^{1/q}[|Y|^q], \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{21.15}$$

For discrete random variables, this was proved in Section 5; see (5.35). For the general case, choose simple random variables X_n and Y_n satisfying $0 \leq |X_n| \uparrow |X|$ and $0 \leq |Y_n| \uparrow |Y|$; see (20.2). Then (5.35) and the monotone convergence theorem give (21.15). Notice that (21.15) implies that if $|X|^p$ and $|Y|^q$ are integrable, then so is XY. Schwarz's inequality is the case $p = q = 2$:

$$E[|XY|] \leq E^{1/2}[X^2] E^{1/2}[Y^2]. \tag{21.16}$$

If X and Y have second moments, then XY must have a first moment.

The same reasoning shows that Lyapounov's inequality (5.37) carries over from the simple to the general case.

Joint Integrals

The relation (21.1) extends to random vectors. Suppose that $(X_1, \ldots, X_k)$ has distribution μ in k-space and $g\colon R^k \to R^1$. By Theorem 16.13,

$$E[g(X_1, \ldots, X_k)] = \int_{R^k} g(x)\mu(dx), \tag{21.17}$$

with the usual provisos about infinite values. For example, $E[X_i X_j] = \int_{R^k} x_i x_j \mu(dx)$. If $E[X_i] = m_i$, the *covariance* of X_i and X_j is

$$\operatorname{Cov}[X_i, X_j] = E[(X_i - m_i)(X_j - m_j)] = \int_{R^k} (x_i - m_i)(x_j - m_j)\mu(dx).$$

Random variables are *uncorrelated* if they have covariance 0.

Independence and Expected Value

Suppose that X and Y are independent. If they are also simple, then $E[XY] = E[X]E[Y]$, as proved in Section 5—see (5.25). Define X_n by (20.2) and similarly define $Y_n = \psi_n(Y^+) - \psi_n(Y^-)$. Then X_n and Y_n are independent and simple, so that $E[|X_n Y_n|] = E[|X_n|]E[|Y_n|]$, and $0 \le |X_n| \uparrow |X|$, $0 \le |Y_n| \uparrow |Y|$. If X and Y are integrable, then $E[|X_n Y_n|] = E[|X_n|]E[|Y_n|] \uparrow E[|X|] \cdot E[|Y|]$, and it follows by the monotone convergence theorem that $E[|XY|] < \infty$; since $X_n Y_n \to XY$ and $|X_n Y_n| \le |XY|$, it follows further by the dominated convergence theorem that $E[XY] = \lim_n E[X_n Y_n] = \lim_n E[X_n]E[Y_n] = E[X]E[Y]$. Therefore, XY is integrable if X and Y are (which is by no means true for dependent random variables) and $E[XY] = E[X]E[Y]$.

This argument obviously extends inductively: If $X_1, \ldots, X_k$ are independent and integrable, then the product $X_1 \cdots X_k$ is also integrable and

$$E[X_1 \cdots X_k] = E[X_1] \cdots E[X_k]. \tag{21.18}$$

Suppose that $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent σ-fields, A lies in $\mathcal{G}_1$, X_1 is measurable $\mathcal{G}_1$, and X_2 is measurable $\mathcal{G}_2$. Then $I_A X_1$ and X_2 are independent, so that (21.18) gives

$$\int_A X_1 X_2 \, dP = \int_A X_1 \, dP \cdot E[X_2] \tag{21.19}$$

if the random variables are integrable. In particular,

$$\int_A X_2 \, dP = P(A) E[X_2]. \tag{21.20}$$

From (21.18) it follows just as for simple random variables (see (5.28)) that variances add for sums of independent random variables. It is even enough that the random variables be independent in pairs.

Moment Generating Functions

The *moment generating function* is defined as

$$M(s) = E\left[e^{sX}\right] = \int_{-\infty}^{\infty} e^{sx}\mu(dx) = \int_{-\infty}^{\infty} e^{sx}\,dF(x) \tag{21.21}$$

for all s for which this is finite (note that the integrand is nonnegative). Section 9 shows in the case of simple random variables the power of moment generating function methods. This function is also called the *Laplace transform* of μ, especially in nonprobabilistic contexts.

Now $\int_0^{\infty} e^{sx}\mu(dx)$ is finite for $s \le 0$, and if it is finite for a positive s, then it is finite for all smaller s. Together with the corresponding result for the left half-line, this shows that $M(s)$ is defined on some interval containing 0. If X is nonnegative, this interval contains $(-\infty, 0]$ and perhaps part of $(0, \infty)$; if X is nonpositive, it contains $[0, \infty)$ and perhaps part of $(-\infty, 0)$. It is possible that the interval consists of 0 alone; this happens, for example, if μ is concentrated on the integers and $\mu\{n\} = \mu\{-n\} = C/n^2$ for $n = 1, 2, \ldots$.

Suppose that $M(s)$ is defined throughout an interval $(-s_0, s_0)$, where $s_0 > 0$. Since $e^{|sx|} \le e^{sx} + e^{-sx}$ and the latter function is integrable μ for $|s| < s_0$, so is $\sum_{k=0}^{\infty} |sx|^k/k! = e^{|sx|}$. By the corollary to Theorem 16.7, μ has finite moments of all orders and

$$M(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} E\left[X^k\right] = \sum_{k=0}^{\infty} \frac{s^k}{k!} \int_{-\infty}^{\infty} x^k \mu(dx), \qquad |s| < s_0. \tag{21.22}$$

Thus $M(s)$ has a Taylor expansion about 0 with positive radius of convergence if it is defined in some $(-s_0, s_0)$, $s_0 > 0$. If $M(s)$ can somehow be calculated and expanded in a series $\sum_k a_k s^k$, and if the coefficients a_k can be identified, then, since a_k must coincide with $E[X^k]/k!$, the moments of X can be computed: $E[X^k] = a_k k!$ It also follows from the theory of Taylor expansions [A29] that $a_k k!$ is the kth derivative $M^{(k)}(s)$ evaluated at $s = 0$:

$$M^{(k)}(0) = E\left[X^k\right] = \int_{-\infty}^{\infty} x^k \mu(dx). \tag{21.23}$$

This holds if $M(s)$ exists in some neighborhood of 0.

Suppose now that M is defined in some neighborhood of s. If ν has density $e^{sx}/M(s)$ with respect to μ (see (16.11)), then ν has moment generating function $N(u) = M(s+u)/M(s)$ for u in some neighborhood of 0.

But then by (21.23), $N^{(k)}(0) = \int_{-\infty}^{\infty} x^k \nu(dx) = \int_{-\infty}^{\infty} x^k e^{sx} \mu(dx)/M(s)$, and since $N^{(k)}(0) = M^{(k)}(s)/M(s)$,

$$M^{(k)}(s) = \int_{-\infty}^{\infty} x^k e^{sx} \mu(dx). \tag{21.24}$$

This holds as long as the moment generating function exists in some neighborhood of s. If $s = 0$, this gives (21.23) again. Taking $k = 2$ shows that $M(s)$ is convex in its interval of definition.

Example 21.2. For the standard normal density,

$$M(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} e^{s^2/2} \int_{-\infty}^{\infty} e^{-(x-s)^2/2}\, dx,$$

and a change of variable gives

$$M(s) = e^{s^2/2}. \tag{21.25}$$

The moment generating function in this case defined for all s. Since $e^{s^2/2}$ has the expansion

$$e^{s^2/2} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{s^2}{2}\right)^k = \sum_{k=0}^{\infty} \frac{1 \times 3 \times \cdots \times (2k-1)}{(2k)!} s^{2k},$$

the moments can be read off from (21.22), which proves (21.7) once more. ■

Example 21.3. In the exponential case (20.10), the moment generating function

$$M(s) = \int_0^{\infty} e^{sx} \alpha e^{-\alpha x}\, dx = \frac{\alpha}{\alpha - s} = \sum_{k=0}^{\infty} \frac{s^k}{\alpha^k} \tag{21.26}$$

is defined for $s < \alpha$. By (21.22) the kth moment is $k!\alpha^{-k}$. The mean and variance are thus α^{-1} and α^{-2}. ■

Example 21.4. For the Poisson distribution (20.7),

$$M(s) = \sum_{r=0}^{\infty} e^{rs} e^{-\lambda} \frac{\lambda^r}{r!} = e^{\lambda(e^s - 1)}. \tag{21.27}$$

Since $M'(s) = \lambda e^s M(s)$ and $M''(s) = (\lambda^2 e^{2s} + \lambda e^s) M(s)$, the first two moments are $M'(0) = \lambda$ and $M''(0) = \lambda^2 + \lambda$; the mean and variance are both λ. ■

Let $X_1, \ldots, X_k$ be independent random variables, and suppose that each X_i has a moment generating function $M_i(s) = E[e^{sX_i}]$ in $(-s_0, s_0)$. For $|s| < s_0$, each $\exp(sX_i)$ is integrable, and, since they are independent, their product $\exp(s\sum_{i=1}^k X_i)$ is also integrable (see (21.18)). The moment generating function of $X_1 + \cdots + X_k$ is therefore

$$M(s) = M_1(s) \cdots M_k(s) \tag{21.28}$$

in $(-s_0, s_0)$. This relation for simple random variables was essential to the arguments in Section 9.

For simple random variables it was shown in Section 9 that the moment generating function determines the distribution. This will later be proved for general random variables; see Theorem 22.2 for the nonnegative case and Section 30 for the general case.

PROBLEMS

21.1. Prove

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-tx^2/2}\, dx = t^{-1/2},$$

differentiate k times with respect to t inside the integral (justify), and derive (21.7) again.

21.2. Show that, if X has the standard normal distribution, then $E[|X|^{2n+1}] = 2^n n! \sqrt{2/\pi}$.

21.3. 20.9↑ *Records.* Consider the sequence of records in the sense of Problem 20.9. Show that the expected waiting time to the next record is infinite.

21.4. 20.14↑ Show that the Cauchy distribution has no mean.

21.5. Prove the first Borel–Cantelli lemma by applying Theorem 16.6 to indicator random variables. Why is Theorem 16.6 not enough for the second Borel–Cantelli lemma?

21.6. Prove (21.9) by Fubini's theorem.

21.7. Prove for integrable X that

$$E[X] = \int_0^{\infty} P[X > t]\, dt - \int_{-\infty}^{0} P[X < t]\, dt.$$

21.8. (a) Suppose that X and Y have first moments, and prove

$$E[Y]-E[X]=\int_{-\infty}^{\infty}(P[X<t\le Y]-P[Y<t\le X])\,dt.$$

(b) Let $(X,Y]$ be a nondegenerate random interval. Show that its expected length is the integral with respect to t of the probability that it covers t.

21.9. Suppose that X and Y are random variables with distribution functions F and G.

(a) Show that if F and G have no common jumps, then $E[F(Y)]+E[G(X)]=1$.

(b) If F is continuous, then $E[F(X)]=\frac{1}{2}$.

(c) Even if F and G have common jumps, if X and Y are taken to be independent, then $E[F(Y)]+E[G(X)]=1+P[X=Y]$.

(d) Even if F has jumps, $E[F(X)]=\frac{1}{2}+\frac{1}{2}\Sigma_x P^2[X=x]$.

21.10. (a) Show that uncorrelated variables need not be independent.

(b) Show that $\operatorname{Var}[\Sigma_{i=1}^n X_i]=\Sigma_{i,j=1}^n \operatorname{Cov}[X_i,X_j]=\Sigma_{i=1}^n \operatorname{Var}[X_i]+2\Sigma_{1\le i<j\le n}\operatorname{Cov}[X_i,X_j]$. The cross terms drop out if the X_i are uncorrelated, and hence drop out if they are independent.

21.11. ↑ Let X, Y, and Z be independent random variables such that X and Y assume the values $0,1,2$ with probability $\frac{1}{3}$ each and Z assumes the values 0 and 1 with probabilities $\frac{1}{3}$ and $\frac{2}{3}$. Let $X'=X$ and $Y'=X+Z$ (mod 3).

(a) Show that X', Y', and $X'+Y'$ have the same one-dimensional distributions as X, Y, and $X+Y$, respectively, even though (X',Y') and (X,Y) have different distributions.

(b) Show that X' and Y' are dependent but uncorrelated.

(c) Show that, despite dependence, the moment generating function of $X'+Y'$ is the product of the moment generating functions of X' and Y'.

21.12. Suppose that X and Y are independent, nonnegative random variables and that $E[X]=\infty$ and $E[Y]=0$. What is the value common to $E[XY]$ and $E[X]E[Y]$? Use the conventions (15.2) for both the product of the random variables and the product of their expected values. What if $E[X]=\infty$ and $0<E[Y]<\infty$?

21.13. Suppose that X and Y are independent and that $f(x,y)$ is nonnegative. Put $g(x)=E[f(x,Y)]$ and show that $E[g(X)]=E[f(X,Y)]$. Show more generally that $\int_{X\in A}g(X)\,dP=\int_{X\in A}f(X,Y)\,dP$. Extend to f that may be negative.

21.14. ↑ The integrability of $X+Y$ does not imply that of X and Y separately. Show that it does if X and Y are independent.

21.15. 20.25 ↑ Write $d_1(X,Y)=E[|X-Y|/(1+|X-Y|)]$. Show that this is a metric equivalent to the one in Problem 20.25.

21.16. For the density $C\exp(-|x|^{1/2})$, $-\infty<x<\infty$, show that moments of all orders exist but that the moment generating function exists only at $s=0$.

21.17. 16.6↑ Show that a moment generating function $M(s)$ defined in $(-s_0, s_0)$, $s_0 > 0$, can be extended to a function analytic in the strip $[z: -s_0 < \operatorname{Re} z < s_0]$. If $M(s)$ is defined in $[0, s_0)$, $s_0 > 0$, show that it can be extended to a function continuous in $[z: 0 \le \operatorname{Re} z < s_0]$ and analytic in $[z: 0 < \operatorname{Re} z < s_0]$.

21.18. Use (21.28) to find the generating function of (20.39).

21.19. For independent random variables having moment generating functions, show by (21.28) that the variances add.

21.20. 20.17↑ Show that the gamma density (20.47) has moment generating function $(1 - s/\alpha)^{-u}$ for $s < \alpha$. Show that the kth moment is $u(u+1)\cdots(u+k-1)/\alpha^k$. Show that the chi-squared distribution with n degrees of freedom has mean n and variance $2n$.

21.21. Let $X_1, X_2, \ldots$ be identically distributed random variables with finite second moment. Show that $nP[|X_1| \ge \epsilon\sqrt{n}\,] \to 0$ and $n^{-1/2} \max_{k \le n} |X_k| \to_P 0$.

SECTION 22. SUMS OF INDEPENDENT RANDOM VARIABLES

Let $X_1, X_2, \ldots$ be a sequence of independent random variables on some probability space. It is natural to ask whether the infinite series $\sum_{n=1}^{\infty} X_n$ converges with probability 1, or as in Section 6 whether $n^{-1}\sum_{k=1}^{n} X_k$ converges to some limit with probability 1. It is to questions of this sort that the present section is devoted.

Throughout the section, S_n will denote the partial sum $\sum_{k=1}^{n} X_k$ $(S_0 = 0)$.

The Strong Law of Large Numbers

The central result is a general version of Theorem 6.1.

Theorem 22.1. *If $X_1, X_2, \ldots$ are independent and identically distributed and have finite mean, then $S_n/n \to E[X_1]$ with probability* 1.

Formerly this theorem stood at the end of a chain of results. The following argument, due to Etemadi, proceeds from first principles.

PROOF. If the theorem holds for nonnegative random variables, then $n^{-1}S_n = n^{-1}\sum_{k=1}^{n} X_k^+ - n^{-1}\sum_{k=1}^{n} X_k^- \to E[X_1^+] - E[X_1^-] = E[X_1]$ with probability 1. Assume then that $X_k \ge 0$.

Consider the truncated random variables $Y_k = X_k I_{[X_k \le k]}$ and their partial sums $S_n^* = \sum_{k=1}^{n} Y_k$. For $\alpha > 1$, temporarily fixed, let $u_n = \lfloor \alpha^n \rfloor$. The first step is to prove

$$\sum_{n=1}^{\infty} P\left[\left|\frac{S_{u_n}^* - E[S_{u_n}^*]}{u_n}\right| > \epsilon\right] < \infty. \tag{22.1}$$

Since the X_n are independent and identically distributed,

$$\begin{aligned}\operatorname{Var}[S_n^*] &= \sum_{k=1}^{n} \operatorname{Var}[Y_k] \le \sum_{k=1}^{n} E[Y_k^2] \\ &= \sum_{k=1}^{n} E\left[X_1^2 I_{[X_1 \le k]}\right] \le nE\left[X_1^2 I_{[X_1 \le n]}\right].\end{aligned}$$

It follows by Chebyshev's inequality that the sum in (22.1) is at most

$$\sum_{n=1}^{\infty} \frac{\operatorname{Var}[S_{u_n}^*]}{\epsilon^2 u_n^2} \le \frac{1}{\epsilon^2} E\left[X_1^2 \sum_{n=1}^{\infty} \frac{1}{u_n} I_{[X_1 \le u_n]}\right].$$

Let $K = 2\alpha/(\alpha - 1)$, and suppose $x > 0$. If N is the smallest n such that $u_n \ge x$, then $\alpha^N \ge x$, and since $y \le 2\lfloor y \rfloor$ for $y \ge 1$,

$$\sum_{u_n \ge x} u_n^{-1} \le 2 \sum_{n \ge N} \alpha^{-n} = K\alpha^{-N} \le Kx^{-1}.$$

Therefore, $\sum_{n=1}^{\infty} u_n^{-1} I_{[X_1 \le u_n]} \le KX_1^{-1}$ for $X_1 > 0$, and the sum in (22.1) is at most $K\epsilon^{-2}E[X_1] < \infty$.

From (22.1) it follows by the first Borel–Cantelli lemma (take a union over positive, rational ϵ) that $(S_{u_n}^* - E[S_{u_n}^*])/u_n \to 0$ with probability 1. But by the consistency of Cesàro summation [A30], $n^{-1}E[S_n^*] = n^{-1}\sum_{k=1}^{n} E[Y_k]$ has the same limit as $E[Y_n]$, namely, $E[X_1]$. Therefore $S_{u_n}^*/u_n \to E[X_1]$ with probability 1. Since

$$\sum_{n=1}^{\infty} P[X_n \ne Y_n] = \sum_{n=1}^{\infty} P[X_1 > n] \le \int_0^{\infty} P[X_1 > t]\, dt = E[X_1] < \infty,$$

another application of the first Borel–Cantelli lemma shows that $(S_n^* - S_n)/n \to 0$ and hence

$$\frac{S_{u_n}}{u_n} \to E[X_1] \tag{22.2}$$

with probability 1.

If $u_n \le k \le u_{n+1}$, then since $X_i \ge 0$,

$$\frac{u_n}{u_{n+1}} \frac{S_{u_n}}{u_n} \le \frac{S_k}{k} \le \frac{u_{n+1}}{u_n} \frac{S_{u_{n+1}}}{u_{n+1}}.$$

But $u_{n+1}/u_n \to \alpha$, and so it follows by (22.2) that

$$\frac{1}{\alpha} E[X_1] \le \liminf_k \frac{S_k}{k} \le \limsup_k \frac{S_k}{k} \le \alpha E[X_1]$$

with probability 1. This is true for each $\alpha > 1$. Intersecting the corresponding sets over rational α exceeding 1 gives $\lim_k S_k/k = E[X_1]$ with probability 1. ■

Although the hypothesis that the X_n all have the same distribution is used several times in this proof, independence is used only through the equation $\mathrm{Var}[S_n^*] = \sum_{k=1}^n \mathrm{Var}[Y_k]$, and for this it is enough that the X_n be independent in pairs. The proof given for Theorem 6.1 of course extends beyond the case of simple random variables, but it requires $E[X_1^4] < \infty$.

Corollary. *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and $E[X_1^-] < \infty$, $E[X_1^+] = \infty$ (so that $E[X_1] = \infty$). Then $n^{-1}\sum_{k=1}^n X_k \to \infty$ with probability* 1.

PROOF. By the theorem, $n^{-1}\sum_{k=1}^n X_k^- \to E[X_1^-]$ with probability 1, and so it suffices to prove the corollary for the case $X_1 = X_1^+ \ge 0$. If

$$X_n^{(u)} = \begin{cases} X_n & \text{if } 0 \le X_n \le u, \\ 0 & \text{if } X_n > u, \end{cases}$$

then $n^{-1}\sum_{k=1}^n X_k \ge n^{-1}\sum_{k=1}^n X_k^{(u)} \to E[X_1^{(u)}]$ by the theorem. Let $u \to \infty$. ■

The Weak Law and Moment Generating Functions

The weak law of large numbers (Section 6) carries over without change to the case of general random variables with second moments—only Chebyshev's inequality is required. The idea can be used to prove in a very simple way that a distribution concentrated on $[0, \infty)$ is uniquely determined by its moment generating function or Laplace transform.

For each λ, let Y_λ be a random variable (on some probability space) having the Poisson distribution with parameter λ. Since Y_λ has mean and variance λ (Example 21.4), Chebyshev's inequality gives

$$P\left[\left|\frac{Y_\lambda - \lambda}{\lambda}\right| \ge \epsilon\right] \le \frac{\lambda}{\lambda^2\epsilon^2} \to 0, \qquad \lambda \to \infty.$$

Let G_λ be the distribution function of Y_λ/λ, so that

$$G_\lambda(t) = \sum_{k=0}^{\lfloor \lambda t \rfloor} e^{-\lambda} \frac{\lambda^k}{k!}.$$

The result above can be restated as

$$\lim_{\lambda \to \infty} G_\lambda(t) = \begin{cases} 1 & \text{if } t > 1, \\ 0 & \text{if } t < 1. \end{cases} \tag{22.3}$$

In the notation of Section 14, $G_\lambda(x) \Rightarrow \Delta(x-1)$ as $\lambda \to \infty$.

Now consider a probability distribution μ concentrated on $[0,\infty)$. Let F be the corresponding distribution function. Define

$$M(s) = \int_0^\infty e^{-sx}\mu(dx), \qquad s \geq 0; \tag{22.4}$$

here 0 is included in the range of integration. This is the moment generating function (21.21), but the argument has been reflected through the origin. It is a *one-sided Laplace transform*, defined for all nonnegative s.

For positive s, (21.24) gives

$$M^{(k)}(s) = (-1)^k \int_0^\infty y^k e^{-sy}\mu(dy). \tag{22.5}$$

Therefore, for positive x and s,

$$\begin{aligned} \sum_{k=0}^{\lfloor sx \rfloor} \frac{(-1)^k}{k!} s^k M^{(k)}(s) &= \int_0^\infty \sum_{k=0}^{\lfloor sx \rfloor} e^{-sy} \frac{(sy)^k}{k!} \mu(dy) \\ &= \int_0^\infty G_{sy}\left(\frac{x}{y}\right)\mu(dy). \end{aligned} \tag{22.6}$$

Fix $x > 0$. If[†] $0 \leq y < x$, then $G_{sy}(x/y) \to 1$ as $s \to \infty$ by (22.3); if $y > x$, the limit is 0. If $\mu\{x\} = 0$, the integrand on the right in (22.6) thus converges as $s \to \infty$ to $I_{[0,x]}(y)$ except on a set of μ-measure 0. The bounded convergence theorem then gives

$$\lim_{s \to \infty} \sum_{k=0}^{\lfloor sx \rfloor} \frac{(-1)^k}{k!} s^k M^{(k)}(s) = \mu[0,x] = F(x). \tag{22.7}$$

[†]If $y = 0$, the integrand in (22.5) is 1 for $k = 0$ and 0 for $k \geq 1$; hence for $y = 0$, the integrand in the middle term of (22.6) is 1.

Thus $M(s)$ determines the value of F at x if $x > 0$ and $\mu\{x\} = 0$, which covers all but countably many values of x in $[0, \infty)$. Since F is right-continuous, F itself and hence μ are determined through (22.7) by $M(s)$. In fact μ is by (22.7) determined by the values of $M(s)$ for s beyond an arbitrary s_0:

Theorem 22.2. *Let μ and ν be probability measures on $[0, \infty)$. If*

$$\int_0^\infty e^{-sx}\mu(dx) = \int_0^\infty e^{-sx}\nu(dx), \qquad s \geq s_0,$$

where $s_0 \geq 0$, then $\mu = \nu$.

Corollary. *Let f_1 and f_2 be real functions on $[0, \infty)$. If*

$$\int_0^\infty e^{-sx} f_1(x)\,dx = \int_0^\infty e^{-sx} f_2(x)\,dx, \qquad s \geq s_0,$$

where $s_0 \geq 0$, then $f_1 = f_2$ outside a set of Lebesgue measure 0.

The f_i need not be nonnegative, and they need not be integrable, but $e^{-sx}f_i(x)$ must be integrable over $[0, \infty)$ for $s \geq s_0$.

PROOF. For the nonnegative case, apply the theorem to the probability densities $g_i(x) = e^{-s_0 x} f_i(x)/m$, where $m = \int_0^\infty e^{-s_0 x} f_i(x)\,dx$, $i = 1, 2$. For the general case, prove that $f_1^+ + f_2^- = f_2^+ + f_1^-$ almost everywhere. ■

Example 22.1. If $\mu_1 * \mu_2 = \mu_3$, then the corresponding transforms (22.4) satisfy $M_1(s)M_2(s) = M_3(s)$ for $s \geq 0$. If μ_i is the Poisson distribution with mean λ_i, then (see (21.27)) $M_i(s) = \exp[\lambda_i(e^{-s} - 1)]$. It follows by Theorem 22.2 that if two of the μ_i are Poisson, so is the third, and $\lambda_1 + \lambda_2 = \lambda_3$. ■

Kolmogorov's Zero–One Law

Consider the set A of ω for which $n^{-1}\sum_{k=1}^n X_k(\omega) \to 0$ as $n \to \infty$. For each m, the values of $X_1(\omega), \ldots, X_{m-1}(\omega)$ are irrelevant to the question of whether or not ω lies in A, and so A ought to lie in the σ-field $\sigma(X_m, X_{m+1}, \ldots)$. In fact, $\lim_n n^{-1}\sum_{k=0}^{m-1} X_k(\omega) = 0$ for fixed m, and hence ω lies in A if and only if $\lim_n n^{-1}\sum_{k=m}^n X_k(\omega) = 0$. Therefore,

$$A = \bigcap_{\epsilon} \bigcup_{N \geq m} \bigcap_{n \geq N} \left[\omega: \left| n^{-1} \sum_{k=m}^{n} X_k(\omega) \right| < \epsilon\right], \tag{22.8}$$

the first intersection extending over positive rational ϵ. The set on the inside

lies in $\sigma(X_m, X_{m+1}, \ldots)$, and hence so does A. Similarly, the ω-set where the series $\sum_n X_n(\omega)$ converges lies in each $\sigma(X_m, X_{m+1}, \ldots)$.

The intersection $\mathscr{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots)$ is the *tail* σ-field associated with the sequence $X_1, X_2, \ldots$; its elements are *tail events*. In the case $X_n = I_{A_n}$, this is the σ-field (4.29) studied in Section 4. The following general form of *Kolmogorov's zero–one law* extends Theorem 4.5.

Theorem 22.3. *Suppose that $\{X_n\}$ is independent and that $A \in \mathscr{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots)$. Then either $P(A) = 0$ or $P(A) = 1$.*

PROOF. Let $\mathscr{F}_0 = \bigcup_{k=1}^{\infty} \sigma(X_1, \ldots, X_k)$. The first thing to establish is that $\mathscr{F}_0$ is a field generating the σ-field $\sigma(X_1, X_2, \ldots)$. If B and C lie in $\mathscr{F}_0$, then $B \in \sigma(X_1, \ldots, X_j)$ and $C \in \sigma(X_1, \ldots, X_k)$ for some j and k; if $m = \max\{j, k\}$, then B and C both lie in $\sigma(X_1, \ldots, X_m)$, so that $B \cup C \in \sigma(X_1, \ldots, X_m) \subset \mathscr{F}_0$. Thus $\mathscr{F}_0$ is closed under the formation of finite unions; since it is similarly closed under complementation, $\mathscr{F}_0$ is a field. For $H \in \mathscr{R}^1$, $[X_n \in H] \in \mathscr{F}_0 \subset \sigma(\mathscr{F}_0)$, and hence X_n is measurable $\sigma(\mathscr{F}_0)$; thus $\mathscr{F}_0$ generates $\sigma(X_1, X_2, \ldots)$ (which in general is much larger than $\mathscr{F}_0$).

Suppose that A lies in $\mathscr{T}$. Then A lies in $\sigma(X_{k+1}, X_{k+2}, \ldots)$ for each k. Therefore, if $B \in \sigma(X_1, \ldots, X_k)$, then A and B are independent by Theorem 20.2. Therefore, A is independent of $\mathscr{F}_0$ and hence by Theorem 4.2 is also independent of $\sigma(X_1, X_2, \ldots)$. But then A is independent of itself: $P(A \cap A) = P(A)P(A)$. Therefore, $P(A) = P^2(A)$, which implies that $P(A)$ is either 0 or 1. ■

As noted above, the set where $\sum_n X_n(\omega)$ converges satisfies the hypothesis of Theorem 22.3, and so does the set where $n^{-1}\sum_{k=1}^{n} X_k(\omega) \to 0$. In many similar cases it is very easy to prove by this theorem that a set at hand must have probability either 0 or 1. But to determine which of 0 and 1 is, in fact, the probability of the set may be extremely difficult.

Maximal Inequalities

Essential to the study of random series are maximal inequalities—inequalities concerning the maxima of partial sums. The best known is that of Kolmogorov.

Theorem 22.4. *Suppose that $X_1, \ldots, X_n$ are independent with mean 0 and finite variances. For $\alpha > 0$,*

$$P\left[\max_{1 \le k \le n} |S_k| \ge \alpha\right] \le \frac{1}{\alpha^2} \operatorname{Var}[S_n]. \tag{22.9}$$

PROOF. Let A_k be the set where $|S_k| \geq \alpha$ but $|S_j| < \alpha$ for $j < k$. Since the A_k are disjoint,

$$\begin{aligned} E[S_n^2] &\geq \sum_{k=1}^{n} \int_{A_k} S_n^2 \, dP \\ &= \sum_{k=1}^{n} \int_{A_k} \left[S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2 \right] dP \\ &\geq \sum_{k=1}^{n} \int_{A_k} \left[S_k^2 + 2S_k(S_n - S_k) \right] dP. \end{aligned}$$

Since A_k and S_k are measurable $\sigma(X_1, \ldots, X_k)$ and $S_n - S_k$ is measurable $\sigma(X_{k+1}, \ldots, X_n)$, and since the means are all 0, it follows by (21.19) and independence that $\int_{A_k} S_k(S_n - S_k)\, dP = 0$. Therefore,

$$\begin{aligned} E[S_n^2] &\geq \sum_{k=1}^{n} \int_{A_k} S_k^2 \, dP \geq \sum_{k=1}^{n} \alpha^2 P(A_k) \\ &= \alpha^2 P\left[\max_{1 \leq k \leq n} |S_k| \geq \alpha \right]. \end{aligned}$$

■

By Chebyshev's inequality, $P[|S_n| \geq \alpha] \leq \alpha^{-2} \operatorname{Var}[S_n]$. That this can be strengthened to (22.9) is an instance of a general phenomenon: For sums of independent variables, if $\max_{k \leq n} |S_k|$ is large, then $|S_n|$ is probably large as well. Theorem 9.6 is an instance of this, and so is the following result, due to Etemadi.

Theorem 22.5. *Suppose that $X_1, \ldots, X_n$ are independent. For $\alpha \geq 0$,*

$$P\left[\max_{1 \leq k \leq n} |S_k| \geq 3\alpha \right] \leq 3 \max_{1 \leq k \leq n} P[|S_k| \geq \alpha]. \tag{22.10}$$

PROOF. Let B_k be the set where $|S_k| \geq 3\alpha$ but $|S_j| < 3\alpha$ for $j < k$. Since the B_k are disjoint,

$$\begin{aligned} P\left[\max_{1 \leq k \leq n} |S_k| \geq 3\alpha \right] &\leq P[|S_n| \geq \alpha] + \sum_{k=1}^{n-1} P(B_k \cap [|S_n| < \alpha]) \\ &\leq P[|S_n| \geq \alpha] + \sum_{k=1}^{n-1} P(B_k \cap [|S_n - S_k| > 2\alpha]) \\ &= P[|S_n| \geq \alpha] + \sum_{k=1}^{n-1} P(B_k) P[|S_n - S_k| > 2\alpha] \\ &\leq P[|S_n| \geq \alpha] + \max_{1 \leq k \leq n} P[|S_n - S_k| \geq 2\alpha] \\ &\leq P[|S_n| \geq \alpha] + \max_{1 \leq k \leq n} \left(P[|S_n| \geq \alpha] + P[|S_k| \geq \alpha] \right) \\ &\leq 3 \max_{1 \leq k \leq n} P[|S_k| \geq \alpha]. \end{aligned}$$

■

If the X_k have mean 0 and Chebyshev's inequality is applied to the right side of (22.10), and if α is replaced by $\alpha/3$, the result is Kolmogorov's inequality (22.9) with an extra factor of 27 on the right side. For this reason, the two inequalities are equally useful for the applications in this section.

Convergence of Random Series

For independent X_n, the probability that $\sum X_n$ converges is either 0 or 1. It is natural to try and characterize the two cases in terms of the distributions of the individual X_n.

Theorem 22.6. *Suppose that $\{X_n\}$ is an independent sequence and $E[X_n] = 0$. If $\sum \mathrm{Var}[X_n] < \infty$, then $\sum X_n$ converges with probability* 1.

PROOF. By (22.9),

$$P\left[\max_{1\le k\le r} |S_{n+k} - S_n| > \epsilon\right] \le \frac{1}{\epsilon^2} \sum_{k=1}^{r} \mathrm{Var}[X_{n+k}].$$

Since the sets on the left are nondecreasing in r, letting $r \to \infty$ gives

$$P\left[\sup_{k\ge 1} |S_{n+k} - S_n| > \epsilon\right] \le \frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \mathrm{Var}[X_{n+k}].$$

Since $\sum \mathrm{Var}[X_n]$ converges,

$$\lim_n P\left[\sup_{k\ge 1} |S_{n+k} - S_n| > \epsilon\right] = 0 \tag{22.11}$$

for each ϵ.

Let $E(n,\epsilon)$ be the set where $\sup_{j,k\ge n} |S_j - S_k| > 2\epsilon$, and put $E(\epsilon) = \bigcap_n E(n,\epsilon)$. Then $E(n,\epsilon) \downarrow E(\epsilon)$, and (22.11) implies $P(E(\epsilon)) = 0$. Now $\bigcup_\epsilon E(\epsilon)$, where the union extends over positive rational ϵ, contains the set where the sequence $\{S_n\}$ is not fundamental (does not have the Cauchy property), and this set therefore has probability 0. ■

Example 22.2. Let $X_n(\omega) = r_n(\omega)a_n$, where the r_n are the Rademacher functions on the unit interval—see (1.13). Then X_n has variance a_n^2, and so $\sum a_n^2 < \infty$ implies that $\sum r_n(\omega)a_n$ converges with probability 1. An interesting special case is $a_n = n^{-1}$. If the signs in $\sum \pm n^{-1}$ are chosen on the toss of a coin, then the series converges with probability 1. The alternating harmonic series $1 - 2^{-1} + 3^{-1} + \cdots$ is thus typical in this respect. ■

If $\sum X_n$ converges with probability 1, then S_n converges with probability 1 to some finite random variable S. By Theorem 20.5, this implies that

$S_n \to_P S$. The reverse implication of course does not hold in general, but it does if the summands are independent.

Theorem 22.7. *For an independent sequence $\{X_n\}$, the S_n converge with probability 1 if and only if they converge in probability.*

PROOF. It is enough to show that if $S_n \to_P S$, then $\{S_n\}$ is fundamental with probability 1. Since

$$P\left[|S_{n+j} - S_n| \geq \epsilon\right] \leq P\left[|S_{n+j} - S| \geq \frac{\epsilon}{2}\right] + P\left[|S_n - S| \geq \frac{\epsilon}{2}\right],$$

$S_n \to_P S$ implies

$$\lim_n \sup_{j \geq 1} P\left[|S_{n+j} - S_n| \geq \epsilon\right] = 0. \tag{22.12}$$

But by (22.10),

$$P\left[\max_{1 \leq j \leq k} |S_{n+j} - S_n| \geq \epsilon\right] \leq 3 \max_{1 \leq j \leq k} P\left[|S_{n+j} - S_n| \geq \frac{\epsilon}{3}\right],$$

and therefore

$$P\left[\sup_{k \geq 1} |S_{n+k} - S_n| > \epsilon\right] \leq 3 \sup_{k \geq 1} P\left[|S_{n+k} - S_n| \geq \frac{\epsilon}{3}\right].$$

It now follows by (22.12) that (22.11) holds, and the proof is completed as before. ■

The final result in this direction, the *three-series theorem*, provides necessary and sufficient conditions for the convergence of $\sum X_n$ in terms of the individual distributions of the X_n. Let $X_n^{(c)}$ be X_n truncated at c: $X_n^{(c)} = X_n I_{[|X_n| \leq c]}$.

Theorem 22.8. *Suppose that $\{X_n\}$ is independent, and consider the three series*

$$\sum P[|X_n| > c], \qquad \sum E[X_n^{(c)}], \qquad \sum \operatorname{Var}[X_n^{(c)}]. \tag{22.13}$$

In order that $\sum X_n$ converge with probability 1 it is necessary that the three series converge for all positive c and sufficient that they converge for some positive c.

PROOF OF SUFFICIENCY. Suppose that the series (22.13) converge, and put $m_n^{(c)} = E[X_n^{(c)}]$. By Theorem 22.6, $\sum(X_n^{(c)} - m_n^{(c)})$ converges with probability 1, and since $\sum m_n^{(c)}$ converges, so does $\sum X_n^{(c)}$. Since $P[X_n \neq X_n^{(c)} \text{ i.o.}] = 0$ by the first Borel–Cantelli lemma, it follows finally that $\sum X_n$ converges with probability 1. ■

Although it is possible to prove necessity in the three-series theorem by the methods of the present section, the simplest and clearest argument uses the central limit theorem as treated in Section 27. This involves no circularity of reasoning, since the three-series theorem is nowhere used in what follows.

PROOF OF NECESSITY. Suppose that $\sum X_n$ converges with probability 1, and fix $c > 0$. Since $X_n \to 0$ with probability 1, it follows that $\sum X_n^{(c)}$ converges with probability 1 and, by the second Borel–Cantelli lemma, that $\sum P[|X_n| > c] < \infty$.

Let $M_n^{(c)}$ and $s_n^{(c)}$ be the mean and standard deviation of $S_n^{(c)} = \sum_{k=1}^n X_k^{(c)}$. If $s_n^{(c)} \to \infty$, then since the $X_n^{(c)} - m_n^{(c)}$ are uniformly bounded, it follows by the central limit theorem (see Example 27.4) that

$$\lim_n P\left[x < \frac{S_n^{(c)} - M_n^{(c)}}{s_n^{(c)}} \leq y\right] = \frac{1}{\sqrt{2\pi}} \int_x^y e^{-t^2/2}\, dt. \tag{22.14}$$

And since $\sum X_n^{(c)}$ converges with probability 1, $s_n^{(c)} \to \infty$ also implies $S_n^{(c)}/s_n^{(c)} \to 0$ with probability 1, so that (Theorem 20.5)

$$\lim_n P\left[|S_n^{(c)}/s_n^{(c)}| \geq \epsilon\right] = 0. \tag{22.15}$$

But (22.14) and (22.15) stand in contradiction: Since

$$P\left[x < \frac{S_n^{(c)} - M_n^{(c)}}{s_n^{(c)}} \leq y, \left|\frac{S_n^{(c)}}{s_n^{(c)}}\right| < \epsilon\right]$$

is greater than or equal to the probability in (22.14) minus that in (22.15), it is positive for all sufficiently large n (if $x < y$). But then

$$x - \epsilon < -M_n^{(c)}/s_n^{(c)} < y + \epsilon,$$

and this cannot hold simultaneously for, say, $(x - \epsilon, y + \epsilon) = (-1, 0)$ and $(x - \epsilon, y + \epsilon) = (0, 1)$. Thus $s_n^{(c)}$ cannot go to ∞, and the third series in (22.13) converges.

And now it follows by Theorem 22.6 that $\sum(X_n^{(c)} - m_n^{(c)})$ converges with probability 1, so that the middle series in (22.13) converges as well. ■

Example 22.3. If $X_n = r_n a_n$, where r_n are the Rademacher functions, then $\sum a_n^2 < \infty$ implies that $\sum X_n$ converges with probability 1. If $\sum X_n$ converges, then a_n is bounded, and for large c the convergence of the third

series in (22.13) implies $\sum a_n^2 < \infty$: If the signs in $\sum \pm a_n$ are chosen on the toss of a coin, then the series converges with probability 1 or 0 according as $\sum a_n^2$ converges or diverges. If $\sum a_n^2$ converges but $\sum |a_n|$ diverges, then $\sum \pm a_n$ is with probability 1 conditionally but not absolutely convergent. ■

Example 22.4. If $a_n \downarrow 0$ but $\sum a_n^2 = \infty$, then $\sum \pm a_n$ converges if the signs are strictly alternating, but diverges with probability 1 if they are chosen on the toss of a coin. ■

Theorems 22.6, 22.7, and 22.8 concern conditional convergence, and in the most interesting cases, $\sum X_n$ converges not because the X_n go to 0 at a high rate but because they tend to cancel each other out. In Example 22.4, the terms cancel well enough for convergence if the signs are strictly alternating, but not if they are chosen on the toss of a coin.

Random Taylor Series*

Consider a power series $\sum \pm z^n$, where the signs are chosen on the toss of a coin. The radius of convergence being 1, the series represents an analytic function in the open unit disk $D_0 = [z: |z| < 1]$ in the complex plane. The question arises whether this function can be extended analytically beyond D_0. The answer is no: With probability 1 the unit circle is the natural boundary.

Theorem 22.9. *Let $\{X_n\}$ be an independent sequence such that*

$$P[X_n = 1] = P[X_n = -1] = \tfrac{1}{2}, \qquad n = 0, 1, \ldots. \tag{22.16}$$

There is probability 0 that

$$F(\omega, z) = \sum_{n=0}^{\infty} X_n(\omega) z^n \tag{22.17}$$

coincides in D_0 with a function analytic in an open set properly containing D_0.

It will be seen in the course of the proof that the ω-set in question lies in $\sigma(X_0, X_1, \ldots)$ and hence has a probability. It is intuitively clear that if the set is measurable at all, it must depend only on the X_n for large n and hence must have probability either 0 or 1.

PROOF. Since

$$|X_n(\omega)| = 1, \qquad n = 0, 1, \ldots \tag{22.18}$$

*This topic, which requires complex variable theory, may be omitted.

with probability 1, the series in (22.17) has radius of convergence 1 outside a set of measure 0.

Consider an open disk $D = [z: |z - \zeta| < r]$, where $\zeta \in D_0$ and $r > 0$. Now (22.17) coincides in D_0 with a function analytic in $D_0 \cup D$ if and only if its expansion

$$F(\omega, z) = \sum_{m=0}^{\infty} \frac{1}{m!} F^{(m)}(\omega, \zeta)(z - \zeta)^m$$

about ζ converges at least for $|z - \zeta| < r$. Let A_D be the set of ω for which this holds. The coefficient

$$a_m(\omega) = \frac{1}{m!} F^{(m)}(\omega, \zeta) = \sum_{n=m}^{\infty} \binom{n}{m} X_n(\omega) \zeta^{n-m}$$

is a complex-valued random variable measurable $\sigma(X_m, X_{m+1}, \ldots)$. By the root test, $\omega \in A_D$ if and only if $\limsup_m |a_m(\omega)|^{1/m} \leq r^{-1}$. For each m_0, the condition for $\omega \in A_D$ can thus be expressed in terms of $a_{m_0}(\omega), a_{m_0+1}(\omega), \ldots$ alone, and so $A_D \in \sigma(X_{m_0}, X_{m_0+1}, \ldots)$. Thus A_D has a probability, and in fact $P(A_D)$ is 0 or 1 by the zero-one law.

Of course, $P(A_D) = 1$ if $D \subset D_0$. The central step in the proof is to show that $P(A_D) = 0$ if D contains points not in D_0. Assume on the contrary that $P(A_D) = 1$ for such a D. Consider that part of the circumference of the unit circle that lies in D, and let k be an integer large enough that this arc has length exceeding $2\pi/k$. Define

$$Y_n(\omega) = \begin{cases} X_n(\omega) & \text{if } n \not\equiv 0 \pmod{k}, \\ -X_n(\omega) & \text{if } n \equiv 0 \pmod{k}. \end{cases}$$

Let B_D be the ω-set where the function

$$G(\omega, z) = \sum_{n=0}^{\infty} Y_n(\omega) z^n \tag{22.19}$$

coincides in D_0 with a function analytic in $D_0 \cup D$.

The sequence $\{Y_0, Y_1, \ldots\}$ has the same structure as the original sequence: the Y_n are independent and assume the values ± 1 with probability $\frac{1}{2}$ each. Since B_D is defined in terms of the Y_n in the same way as A_D is defined in terms of the X_n, it is intuitively clear that $P(B_D)$ and $P(A_D)$ must be the same. Assume for the moment the truth of this statement, which is somewhat more obvious than its proof.

If for a particular ω each of (22.17) and (22.19) coincides in D_0 with a function analytic in $D_0 \cup D$, the same must be true of

$$F(\omega, z) - G(\omega, z) = 2 \sum_{m=0}^{\infty} X_{mk}(\omega) z^{mk}. \tag{22.20}$$

Let $D_l = [ze^{2\pi il/k}: z \in D]$. Since replacing z by $ze^{2\pi i/k}$ leaves the function (22.20) unchanged, it can be extended analytically to each $D_0 \cup D_l$, $l = 1, 2, \ldots$. Because of the choice of k, it can therefore be extended analytically to $[z: |z| < 1 + \epsilon]$ for some positive ϵ; but this is impossible if (22.18) holds, since the radius of convergence must then be 1.

Therefore, $A_D \cap B_D$ cannot contain a point ω satisfying (22.18). Since (22.18) holds with probability 1, this rules out the possibility $P(A_D) = P(B_D) = 1$ and by the zero–one law leaves only the possibility $P(A_D) = P(B_D) = 0$. Let A be the ω-set where (22.17) extends to a function analytic in some open set larger than D_0. Then $\omega \in A$ if and only if (22.17) extends to $D_0 \cup D$ for some $D = [z: |z - \zeta| < r]$ for which $D - D_0 \neq \varnothing$, r is rational, and ζ has rational real and imaginary parts; in other words, A is the countable union of A_D for such D. Therefore, A lies in $\sigma(X_0, X_1, \ldots)$ and has probability 0.

It remains only to show that $P(A_D) = P(B_D)$, and this is most easily done by comparing $\{X_n\}$ and $\{Y_n\}$ with a canonical sequence having the same structure. Put $Z_n(\omega) = (X_n(\omega) + 1)/2$, and let $T\omega$ be $\sum_{n=0}^{\infty} Z_n(\omega) 2^{-n-1}$ on the ω-set A^* where this sum lies in $(0, 1]$; on $\Omega - A^*$ let $T\omega$ be 1, say. Because of (22.16) $P(A^*) = 1$. Let $\mathscr{F} = \sigma(X_0, X_1, \ldots)$ and let $\mathscr{B}$ be the σ-field of Borel subsets of $(0, 1]$; then $T: \Omega \to (0, 1]$ is measurable $\mathscr{F}/\mathscr{B}$. Let $r_n(x)$ be the nth Rademacher function. If $M = [x: r_i(x) = u_i,\ i = 1, \ldots, n]$, where $u_i = \pm 1$ for each i, then $P(T^{-1}M) = P[\omega: X_i(\omega) = u_i,\ i = 0, 1, \ldots, n - 1] = 2^{-n}$, which is the Lebesgue measure $\lambda(M)$ of M. Since these sets form a π-system generating $\mathscr{B}$, $P(T^{-1}M) = \lambda(M)$ for all M in $\mathscr{B}$ (Theorem 3.3).

Let M_D be the set of x for which $\sum_{n=0}^{\infty} r_{n+1}(x) z^n$ extends analytically to $D_0 \cup D$. Then M_D lies in $\mathscr{B}$, this being a special case of the fact that A_D lies in $\mathscr{F}$. Moreover, if $\omega \in A^*$, then $\omega \in A_D$ if and only if $T\omega \in M_D$: $A^* \cap A_D = A^* \cap T^{-1}M_D$. Since $P(A^*) = 1$, it follows that $P(A_D) = \lambda(M_D)$.

This argument only uses (22.16), and therefore it applies to $\{Y_n\}$ and B_D as well. Therefore, $P(B_D) = \lambda(M_D) = P(A_D)$. ■

PROBLEMS

22.1. Suppose that $X_1, X_2, \ldots$ is an independent sequence and Y is measurable $\sigma(X_n, X_{n+1}, \ldots)$ for each n. Show that there exists a constant a such that $P[Y = a] = 1$.

22.2. Assume $\{X_n\}$ independent, and define $X_n^{(c)}$ as in Theorem 22.8. Prove that for $\sum |X_n|$ to converge with probability 1 it is necessary that $\sum P[|X_n| > c]$ and $\sum E[|X_n^{(c)}|]$ converge for all positive c and sufficient that they converge for some positive c. If the three series (22.13) converge but $\sum E[|X_n^{(c)}|] = \infty$, then there is probability 1 that $\sum X_n$ converges conditionally but not absolutely.

22.3. ↑ (a) Generalize the Borel–Cantelli lemmas: Suppose X_n are nonnegative If $\sum E[X_n] < \infty$, then $\sum X_n$ converges with probability 1. If the X_n are independent and uniformly bounded, and if $\sum E[X_n] = \infty$, then $\sum X_n$ diverges with probability 1.

(b) Construct independent, nonnegative X_n such that $\sum X_n$ converges with probability 1 but $\sum E[X_n]$ diverges. For an extreme example, arrange that $P[X_n > 0 \text{ i.o.}] = 0$ but $E[X_n] \equiv \infty$.

22.4. Show under the hypothesis of Theorem 22.6 that $\sum X_n$ has finite variance and extend Theorem 22.4 to infinite sequences.

22.5. 20.14 22.1↑ Suppose that $X_1, X_2, \ldots$ are independent, each with the Cauchy distribution (20.45) for a common value of u.

(a) Show that $n^{-1}\sum_{k=1}^{n} X_k$ does not converge with probability 1. Contrast with Theorem 22.1.

(b) Show that $P[n^{-1} \max_{k \le n} X_k \le x] \to e^{-u/\pi x}$ for $x > 0$. Relate to Theorem 14.3.

22.6. If $X_1, X_2, \ldots$ are independent and identically distributed, and if $P[X_1 \ge 0] = 1$ and $P[X_1 > 0] > 0$, then $\sum_n X_n = \infty$ with probability 1. Deduce this from Theorem 22.1 and its corollary and also directly: find a positive ϵ such that $X_n > \epsilon$ infinitely often with probability 1.

22.7. Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and $E[|X_1|] = \infty$. Use (21.9) to show that $\sum_n P[|X_n| \ge an] = \infty$ for each a, and conclude that $\sup_n n^{-1}|X_n| = \infty$ with probability 1. Now show that $\sup_n n^{-1}|S_n| = \infty$ with probability 1. Compare with the corollary to Theorem 22.1.

22.8. *Wald's equation.* Let $X_1, X_2, \ldots$ be independent and identically distributed with finite mean, and put $S_n = X_1 + \cdots + X_n$. Suppose that τ is a stopping time: τ has positive integers as values and $[\tau = n] \in \sigma(X_1, \ldots, X_n)$; see Section 7 for examples. Suppose also that $E[\tau] < \infty$.

(a) Prove that

$$E[S_\tau] = E[X_1]E[\tau]. \tag{22.21}$$

(b) Suppose that X_n is ± 1 with probabilities p and q, $p \ne q$, let τ be the first n for which S_n is $-a$ or b (a and b positive integers), and calculate $E[\tau]$. This gives the expected duration of the game in the gambler's ruin problem for unequal p and q.

22.9. 20.9↑ Let Z_n be 1 or 0 according as at time n there is or is not a record in the sense of Problem 20.9. Let $R_n = Z_1 + \cdots + Z_n$ be the number of records up to time n. Show that $R_n/\log n \to_P 1$.

22.10. 22.1↑ (a) Show that for an independent sequence $\{X_n\}$ the radius of convergence of the random Taylor series $\sum_n X_n z^n$ is r with probability 1 for some nonrandom r.

(b) Suppose that the X_n have the same distribution and $P[X_1 \ne 0] > 0$. Show that r is 1 or 0 according as $\log^+|X_1|$ has finite mean or not.

22.11. Suppose that $X_0, X_1, \ldots$ are independent and each is uniformly distributed over $[0, 2\pi]$. Show that with probability 1 the series $\sum_n e^{iX_n} z^n$ has the unit circle as its natural boundary.

22.12. Prove (what is essentially Kolmogorov's zero–one law) that if A is independent of a π-system $\mathscr{P}$ and $A \in \sigma(\mathscr{P})$, then $P(A)$ is either 0 or 1.

22.13. Suppose that $\mathscr{A}$ is a semiring containing Ω.

(a) Show that if $P(A \cap B) \le bP(B)$ for all $B \in \mathscr{A}$, and if $b < 1$ and $A \in \sigma(\mathscr{A})$, then $P(A) = 0$.

(b) Show that if $P(A \cap B) \le P(A)P(B)$ for all $B \in \mathscr{A}$, and if $A \in \sigma(\mathscr{A})$, then $P(A)$ is 0 or 1.

(c) Show that if $aP(B) \le P(A \cap B)$ for all $B \in \mathscr{A}$, and if $a > 0$ and $A \in \sigma(\mathscr{A})$, then $P(A) = 1$.

(d) Show that if $P(A)P(B) \le P(A \cap B)$ for all $B \in \mathscr{A}$, and if $A \in \sigma(\mathscr{A})$, then $P(A)$ is 0 or 1.

(e) Reconsider Problem 3.20.

22.14. 22.12↑ *Burstin's theorem.* Let f be a Borel function on $[0, 1]$ with arbitrarily small periods: For each ϵ there is a p such that $0 < p < \epsilon$ and $f(x) = f(x + p)$ for $0 \le x \le 1 - p$. Show that such an f is constant almost everywhere:

(a) Show that it is enough to prove that $P(f^{-1}B)$ is 0 or 1 for every Borel set B, where P is Lebesgue measure on the unit interval.

(b) Show that $f^{-1}B$ is independent of each interval $[0, x]$, and conclude that $P(f^{-1}B)$ is 0 or 1.

(c) Show by example that f need not be constant.

22.15. Assume that $X_1, \ldots, X_n$ are independent and s, t, α are nonnegative. Let

$$L(s) = \max_{k \le n} P[|S_k| \ge s], \quad R(s) = \max_{k \le n} P[|S_n - S_k| > s],$$

$$M(s) = P\Big[\max_{k \le n} |S_k| \ge s\Big], \quad T(s) = P[|S_n| \ge s].$$

(a) Following the first part of the proof of (22.10), show that

$$M(s + t) \le T(t) + M(s + t)R(s). \tag{22.22}$$

(b) Take $s = 2\alpha$ and $t = \alpha$; use (22.22), together with the inequalities $T(s) \le L(s)$ and $R(2s) \le 2L(s)$, to prove Etemadi's inequality (22.10) in the form

$$M(3\alpha) \le B_E(\alpha) = 1 \wedge 3L(\alpha). \tag{22.23}$$

(c) Carry the rightmost term in (22.22) to the left side, take $s = t = \alpha$, and prove *Ottaviani's* inequality:

$$M(2\alpha) \le B_O(\alpha) = 1 \wedge \frac{T(\alpha)}{1 - R(\alpha)}. \tag{22.24}$$

(d) Prove

$$B_E(\alpha) \le 3B_O(\alpha/2), \quad B_O(\alpha) \le 3B_E(\alpha/6).$$

This shows that the Etemadi and Ottaviani inequalities are of the same power for most purposes (as, for example, for the proofs of Theorem 22.7 and (37.9)). Etemadi's inequality seems the more natural of the two. Neither inequality can replace (9.39) in the proof of the law of the iterated logarithm.

SECTION 23. THE POISSON PROCESS

Characterization of the Exponential Distribution

Suppose that X has the exponential distribution with parameter α:

$$P[X > x] = e^{-\alpha x}, \qquad x \geq 0. \tag{23.1}$$

The definition (4.1) of conditional probability then gives

$$P[X > x + y | X > x] = P[X > y], \qquad x, y \geq 0. \tag{23.2}$$

Image X as the waiting time for the occurrence of some event such as the arrival of the next customer at a queue or telephone call at an exchange. As observed in Section 14 (see (14.6)), (23.2) attributes to the waiting-time mechanism a lack of memory or aftereffect. And as shown in Section 14, the condition (23.2) implies that X has the distribution (23.1) for some positive α. Thus if in the sense of (23.2) there is no aftereffect in the waiting-time mechanism, then the waiting time itself necessarily follows the exponential law.

The Poisson Process

Consider next a stream or sequence of events, say arrivals of calls at an exchange. Let X_1 be the waiting time to the first event, let X_2 be the waiting time between the first and second events, and so on. The formal model consists of an infinite sequence $X_1, X_2, \ldots$ of random variables on some probability space, and $S_n = X_1 + \cdots + X_n$ represents the time of occurrence of the nth event; it is convenient to write $S_0 = 0$. The stream of events itself remains intuitive and unformalized, and the mathematical definitions and arguments are framed in terms of the X_n.

If no two of the events are to occur simultaneously, the S_n must be strictly increasing, and if only finitely many of the events are to occur in each finite interval of time, S_n must go to infinity:

$$0 = S_0(\omega) < S_1(\omega) < S_2(\omega) < \cdots, \qquad \sup_n S_n(\omega) = \infty. \tag{23.3}$$

This condition is the same thing as

$$X_1(\omega) > 0, \quad X_2(\omega) > 0, \ldots, \qquad \sum_n X_n(\omega) = \infty. \tag{23.4}$$

Throughout the section it will be assumed that these conditions hold everywhere—for every ω. If they hold only on a set A of probability 1, and if $X_n(\omega)$ is redefined as $X_n(\omega) = 1$, say, for $\omega \notin A$, then the conditions hold everywhere and the joint distributions of the X_n and S_n are unaffected.

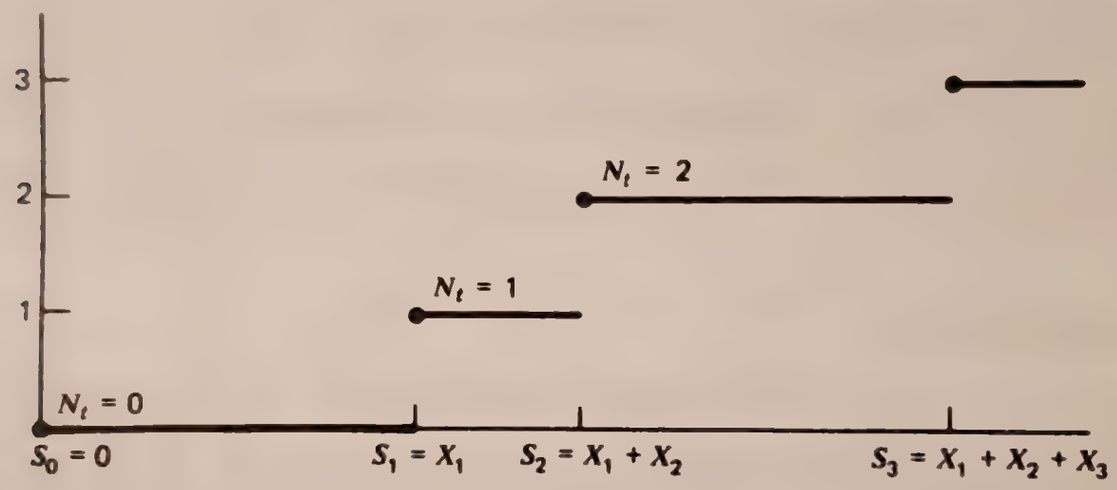
Condition 0°. For each ω, (23.3) *and* (23.4) *hold.*

The arguments go through under the weaker condition that (23.3) and (23.4) hold with probability 1, but they then involve some fussy and uninteresting details. There are at the outset no further restrictions on the X_i; they are not assumed independent, for example, or identically distributed.

The number N_t of events that occur in the time interval $[0, t]$ is the largest integer n such that $S_n \le t$:

$$N_t = \max[n: S_n \le t]. \tag{23.5}$$

Note that $N_t = 0$ if $t < S_1 = X_1$; in particular, $N_0 = 0$. The number of events in $(s, t]$ is the increment $N_t - N_s$.



From (23.5) follows the basic relation connecting the N_t with the S_n:

$$[N_t \ge n] = [S_n \le t]. \tag{23.6}$$

From this follows

$$[N_t = n] = [S_n \le t < S_{n+1}]. \tag{23.7}$$

Each N_t is thus a random variable.

The collection $[N_t: t \ge 0]$ is a *stochastic process*, that is, a collection of random variables indexed by a parameter regarded as time. Condition 0° can be restated in terms of this process:

Condition 0°. For each ω, $N_t(\omega)$ *is a nonnegative integer for* $t \ge 0$, $N_0(\omega) = 0$, *and* $\lim_{t\to\infty} N_t(\omega) = \infty$; *further, for each* ω, $N_t(\omega)$ *as a function of* t *is nondecreasing and right-continuous, and at the points of discontinuity the saltus* $N_t(\omega) - \sup_{s<t} N_s(\omega)$ *is exactly* 1.

It is easy to see that (23.3) and (23.4) and the definition (23.5) give random variables N_t having these properties. On the other hand, if the stochastic

process $[N_t: t \geq 0]$ is given and does have these properties, and if random variables are defined by $S_n(\omega) = \inf[t: N_t(\omega) \geq n]$ and $X_n(\omega) = S_n(\omega) - S_{n-1}(\omega)$, then (23.3) and (23.4) hold, and the definition (23.5) gives back the original N_t. Therefore, anything that can be said about the X_n can be stated in terms of the N_t, and conversely. The points $S_1(\omega), S_2(\omega), \ldots$ of $(0, \infty)$ are exactly the discontinuities of $N_t(\omega)$ as a function of t; because of the queueing example, it is natural to call them *arrival times*.

The program is to study the joint distributions of the N_t under conditions on the waiting times X_n and vice versa. The most common model specifies the independence of the waiting times and the absence of aftereffect:

Condition 1°. The X_n are independent, and each is exponentially distributed with parameter α.

In this case $P[X_n > 0] = 1$ for each n and $n^{-1}S_n \to \alpha^{-1}$ by the strong law of large numbers (Theorem 22.1), and so (23.3) and (23.4) hold with probability 1; to assume they hold everywhere (Condition 0°) is simply a convenient normalization.

Under Condition 1°, S_n has the distribution function specified by (20.40), so that $P[N_t \geq n] = \sum_{i=n}^{\infty} e^{-\alpha t}(\alpha t)^i / i!$ by (23.6), and

$$P[N_t = n] = e^{-\alpha t}\frac{(\alpha t)^n}{n!}, \qquad n = 0, 1, \ldots. \tag{23.8}$$

Thus N_t has the Poisson distribution with mean αt. More will be proved in a moment.

Condition 2°. (i) *For $0 < t_1 < \cdots < t_k$ the increments $N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}}$ are independent.*

(ii) *The individual increments have the Poisson distribution*:

$$P[N_t - N_s = n] = e^{-\alpha(t-s)}\frac{(\alpha(t-s))^n}{n!}, \qquad n = 0, 1, \ldots, \quad 0 \leq s < t. \tag{23.9}$$

Since $N_0 = 0$, (23.8) is a special case of (23.9). A collection $[N_t: t \geq 0]$ of random variables satisfying Condition 2° is called a *Poisson process*, and α is the *rate* of the process. As the increments are independent by (i), if $r < s < t$, then the distributions of $N_s - N_r$ and $N_t - N_s$ must convolve to that of $N_t - N_r$. But the requirement is consistent with (ii) because Poisson distributions with parameters u and v convolve to a Poisson distribution with parameter $u + v$.

Theorem 23.1. *Conditions 1° and 2° are equivalent in the presence of Condition 0°.*

PROOF OF $1^\circ \to 2^\circ$. Fix t, and consider the events that happen after time t. By (23.5), $S_{N_t} \le t < S_{N_t+1}$, and the waiting time from t to the first event following t is $S_{N_t+1} - t$; the waiting time between the first and second events following t is X_{N_t+2}; and so on. Thus

$$(23.10) \qquad X_1^{(t)} = S_{N_t+1} - t, \qquad X_2^{(t)} = X_{N_t+2}, \qquad X_3^{(t)} = X_{N_t+3}, \cdots$$

define the waiting times following t. By (23.6), $N_{t+s} - N_t \ge m$, or $N_{t+s} \ge N_t + m$, if and only if $S_{N_t+m} \le t+s$, which is the same thing as $X_1^{(t)} + \cdots + X_m^{(t)} \le s$. Thus

$$(23.11) \qquad N_{t+s} - N_t = \max\bigl[m\colon X_1^{(t)} + \cdots + X_m^{(t)} \le s\bigr].$$

Hence $[N_{t+s} - N_t = m] = [X_1^{(t)} + \cdots + X_m^{(t)} \le s < X_1^{(t)} + \cdots + X_{m+1}^{(t)}]$. A comparison of (23.11) and (23.5) shows that for fixed t the random variables $N_{t+s} - N_t$ for $s \ge 0$ are defined in terms of the sequence (23.10) in exactly the same way as the N_s are defined in terms of the original sequence of waiting times.

The idea now is to show that conditionally on the event $[N_t = n]$ the random variables (23.10) are independent and exponentially distributed. Because of the independence of the X_k and the basic property (23.2) of the exponential distribution, this seems intuitively clear. For a proof, apply (20.30). Suppose $y \ge 0$; if G_n is the distribution function of S_n, then since X_{n+1} has the exponential distribution,

$$\begin{aligned} P[S_n \le t < S_{n+1},\, S_{n+1} - t > y] &= P[S_n \le t,\, X_{n+1} > t + y - S_n] \\ &= \int_{x \le t} P[X_{n+1} > t + y - x]\, dG_n(x) \\ &= e^{-\alpha y} \int_{x \le t} P[X_{n+1} > t - x]\, dG_n(x) \\ &= e^{-\alpha y} P[S_n \le t,\, X_{n+1} > t - S_n]. \end{aligned}$$

By the assumed independence of the X_n,

$$\begin{aligned} &P\bigl[S_{n+1} - t > y_1,\, X_{n+2} > y_2, \ldots, X_{n+j} > y_j,\, S_n \le t < S_{n+1}\bigr] \\ &\quad = P[S_{n+1} - t > y_1,\, S_n \le t < S_{n+1}]\, e^{-\alpha y_2} \cdots e^{-\alpha y_j} \\ &\quad = P[S_n \le t < S_{n+1}]\, e^{-\alpha y_1} \cdots e^{-\alpha y_j}. \end{aligned}$$

If $H = (y_1, \infty) \times \cdots \times (y_j, \infty)$, this is

$$(23.12) \qquad P\bigl[N_t = n,\, \bigl(X_1^{(t)}, \ldots, X_j^{(t)}\bigr) \in H\bigr] = P[N_t = n]\, P\bigl[(X_1, \ldots, X_j) \in H\bigr].$$

By Theorem 10.4, the equation extends from H of the special form above to all H in $\mathscr{R}^j$.

Now the event $[N_{s_i} = m_i,\ 1 \le i \le u]$ can be put in the form $[(X_1, \ldots, X_j) \in H]$, where $j = m_u + 1$ and H is the set of x in R^j for which $x_1 + \cdots + x_{m_i} \le s_i < x_1 + \cdots + x_{m_i+1}$, $1 \le i \le u$. But then $[(X_1^{(t)}, \ldots, X_j^{(t)}) \in H]$ is by (23.11) the same as the event $[N_{t+s_i} - N_t = m_i,\ 1 \le i \le u]$. Thus (23.12) gives

$$P\left[N_t = n,\ N_{t+s_i} - N_t = m_i,\ 1 \le i \le u\right] = P[N_t = n]P\left[N_{s_i} = m_i,\ 1 \le i \le u\right].$$

From this it follows by induction on k that if $0 = t_0 < t_1 < \cdots < t_k$, then

$$P\left[N_{t_i} - N_{t_{i-1}} = n_i,\ 1 \le i \le k\right] = \prod_{i=1}^{k} P\left[N_{t_i - t_{i-1}} = n_i\right]. \tag{23.13}$$

Thus Condition 1° implies (23.13) and, as already seen, (23.8). But from (23.13) and (23.8) follow the two parts of Condition 2°. ■

PROOF OF 2° → 1°. If 2° holds, then by (23.6), $P[X_1 > t] = P[N_t = 0] = e^{-\alpha t}$, so that X_1 is exponentially distributed. To find the joint distribution of X_1 and X_2, suppose that $0 \le s_1 < t_1 < s_2 < t_2$ and perform the calculation

$$\begin{aligned}
&P[s_1 < S_1 \le t_1,\ s_2 < S_2 < t_2] \\
&\quad = P\left[N_{s_1} = 0,\ N_{t_1} - N_{s_1} = 1,\ N_{s_2} - N_{t_1} = 0,\ N_{t_2} - N_{s_s} \ge 1\right] \\
&\quad = e^{-\alpha s_1} \times \alpha(t_1 - s_1)e^{-\alpha(t_1 - s_1)} \times e^{-\alpha(s_2 - t_1)} \times \left(1 - e^{-\alpha(t_2 - s_2)}\right) \\
&\quad = \alpha(t_1 - s_1)(e^{-\alpha s_2} - e^{-\alpha t_2}) = \iint_{\substack{s_1 < y_1 \le t_1 \\ s_2 < y_2 \le t_2}} \alpha^2 e^{-\alpha y_2}\, dy_1\, dy_2.
\end{aligned}$$

Thus for a rectangle A contained in the open set $G = [(y_1, y_2)\colon 0 < y_1 < y_2]$,

$$P[(S_1, S_2) \in A] = \int_A \alpha^2 e^{-\alpha y_2}\, dy_1\, dy_2.$$

By inclusion–exclusion, this holds for finite unions of such rectangles and hence, by a passage to the limit, for countable ones. Therefore, it holds for $A = G \cap G'$ if G' is open. Since the open sets form a π-system generating the Borel sets, (S_1, S_2) has density $\alpha^2 e^{-\alpha y_2}$ on G (of course, the density is 0 outside G).

By a similar argument in R^k (the notation only is more complicated), $(S_1, \ldots, S_k)$ has density $\alpha^k e^{-\alpha y_k}$ on $[y\colon 0 < y_1 < \cdots < y_k]$. If a linear transformation $g(y) = x$ is defined by $x_i = y_i - y_{i-1}$, then $(X_1, \ldots; X_k) = g(S_1, \ldots, S_k)$ has by (20.20) the density $\prod_{i=1}^{k} \alpha e^{-\alpha x_i}$ (the Jacobian is identically 1). This proves Condition 1°. ■

The Poisson Approximation

Other characterizations of the Poisson process depend on a generalization of the classical Poisson approximation to the binomial distribution.

Theorem 23.2. *Suppose that for each n, $Z_{n1}, \ldots, Z_{nr_n}$ are independent random variables and Z_{nk} assumes the values 1 and 0 with probabilities p_{nk} and $1 - p_{nk}$. If*

$$\sum_{k=1}^{r_n} p_{nk} \to \lambda \geq 0, \qquad \max_{1 \leq k \leq r_n} p_{nk} \to 0, \tag{23.14}$$

then

$$P\left[\sum_{k=1}^{r_n} Z_{nk} = i\right] \to e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, 2, \ldots. \tag{23.15}$$

If $\lambda = 0$, the limit in (23.15) is interpreted as 1 for $i = 0$ and 0 for $i \geq 1$. In the case where $r_n = n$ and $p_{nk} = \lambda/n$, (23.15) is the Poisson approximation to the binomial. Note that if $\lambda > 0$, then (23.14) implies $r_n \to \infty$.

PROOF. The argument depends on a construction like that in the proof of Theorem 20.4. Let $U_1, U_2, \ldots$ be independent random variables, each uniformly distributed over $[0, 1)$. For each p, $0 \leq p \leq 1$, split $[0, 1)$ into the two intervals $I_0(p) = [0, 1 - p)$ and $I_1(p) = [1 - p, 1)$, as well as into the sequence of intervals $J_i(p) = [\sum_{j=1}^{i-1} e^{-p} p^j/j!, \sum_{j=1}^{i} e^{-p} p^j/j!)$, $i = 0, 1, \ldots$. Define $V_{nk} = 1$ if $U_k \in I_1(p_{nk})$ and $V_{nk} = 0$ if $U_k \in I_0(p_{nk})$. Then $V_{n1}, \ldots, V_{nr_n}$ are independent, and V_{nk} assumes the values 1 and 0 with probabilities $P[U_k \in I_1(p_{nk})] = p_{nk}$ and $P[U_k \in I_0(p_{Nk})] = 1 - p_{nk}$. Since $V_{n1}, \ldots, V_{nr_n}$ have the same joint distribution as $Z_{n1}, \ldots, Z_{nr_n}$, (23.15) will follow if it is shown that $V_n = \sum_{k=1}^{r_n} V_{nk}$ satisfies

$$P[V_n = i] \to e^{-\lambda} \frac{\lambda^i}{i!}. \tag{23.16}$$

Now define $W_{nk} = i$ if $U_k \in J_i(p_{nk})$, $i = 0, 1, \ldots$. Then $P[W_{nk} = i] = e^{-p_{nk}} p_{nk}^i/i!$—$W_{nk}$ has the Poisson distribution with mean p_{nk}. Since the W_{nk}

are independent, $W_n = \sum_{k=1}^{r_n} W_{nk}$ has the Poisson distribution with mean $\lambda_n = \sum_{k=1}^{r_n} p_{nk}$. Since $1 - p \le e^{-p}$, $J_1(p) \subset I_1(p)$ (see the diagram). Therefore,

$$P[V_{nk} \ne W_{nk}] = P[V_{nk} = 1 \ne W_{nk}] = P[U_k \in I_1(p_{nk}) - J_1(p_{nk})]$$
$$= p_{nk} - e^{-p_{nk}} p_{nk} \le p_{nk}^2,$$

and

$$P[V_n \ne W_n] \le \sum_{k=1}^{r_n} p_{nk}^2 \le \lambda_n \max_{1 \le k \le r_n} p_{nk} \to 0$$

by (23.14). And now (23.16) and (23.15) follow because

$$P[W_n = i] = e^{-\lambda_n} \lambda_n^i / i! \to e^{-\lambda} \lambda^i / i! \qquad \blacksquare$$

Other Characterizations of the Poisson Process

The condition (23.2) is an interesting characterization of the exponential distribution because it is essentially qualitative. There are qualitative characterizations of the Poisson process as well.

For each ω, the function $N_t(\omega)$ has a discontinuity at t if and only if $S_n(\omega) = t$ for some $n \ge 1$; t is a *fixed discontinuity* if the probability of this is positive. The condition that there be no fixed discontinuities is therefore

$$P[S_n = t] = 0, \qquad t \ge 0, \quad n \ge 1; \tag{23.17}$$

that is, each of $S_1, S_2, \ldots$ has a continuous distribution function. Of course there is probability 1 (under Condition 0°) that $N_t(\omega)$ has a discontinuity *somewhere* (and indeed has infinitely many of them). But (23.17) ensures that a *t specified in advance* has probability 0 of being a discontinuity, or time of an arrival. The Poisson process satisfies this natural condition.

Theorem 23.3. *If Condition 0° holds and* $[N_t: t \ge 0]$ *has independent increments and no fixed discontinuities, then each increment has a Poisson distribution.*

This is Prékopà's theorem. The conclusion is not that $[N_t: t \ge 0]$ is a Poisson process, because the mean of $N_t - N_s$ need not be proportional to $t - s$. If φ is an arbitrary nondecreasing, continuous function on $[0, \infty)$ and $\varphi(0) = 0$, and if $[N_t: t \ge 0]$ is a Poisson process, then $N_{\varphi(t)}$ satisfies the conditions of the theorem.†

PROOF. The problem is to show for $t' < t''$ that $N_{t''} - N_{t'}$ has for some $\lambda \ge 0$ a Poisson distribution with mean λ, a unit mass at 0 being regarded as a Poisson distribution with mean 0.

†This is in fact the general process satisfying them; see Problem 23.8.

The procedure is to construct a sequence of partitions

$$t' = t_{n0} < t_{n1} < \cdots < t_{nr_n} = t'' \tag{23.18}$$

of $[t', t'']$ with three properties. First, each decomposition refines the preceding one: each t_{nk} is a $t_{n+1,j}$. Second,

$$\sum_{k=1}^{r_n} P\left[N_{t_{nk}} - N_{t_{n,k-1}} \geq 1\right] \uparrow \lambda \tag{23.19}$$

for some finite λ and

$$\max_{1 \leq k \leq r_n} P\left[N_{t_{nk}} - N_{t_{n,k-1}} \geq 1\right] \to 0. \tag{23.20}$$

Third,

$$P\left[\max_{1 \leq k \leq r_n} \left(N_{t_{nk}} - N_{t_{n,k-1}}\right) \geq 2\right] \to 0. \tag{23.21}$$

Once the partitions have been constructed, the rest of the proof is easy: Let Z_{nk} be 1 or 0 according as $N_{t_{nk}} - N_{t_{n,k-1}}$ is positive or not. Since $[N_t: t \geq 0]$ has independent increments, the Z_{nk} are independent for each n. By Theorem 23.2, therefore, (23.19) and (23.20) imply that $Z_n = \sum_{k=1}^{r_n} Z_{nk}$ satisfies $P[Z_n = i] \to e^{-\lambda}\lambda^i/i!$ Now $N_{t''} - N_{t'} \geq Z_n$, and there is strict inequality if and only if $N_{t_{nk}} - N_{t_{n,k-1}} \geq 2$ for some k. Thus (23.21) implies $P[N_{t''} - N_{t'} \neq Z_n] \to 0$, and therefore $P[N_{t''} - N_{t'} = i] = e^{-\lambda}\lambda^i/i!$

To construct the partitions, consider for each t the distance $D_t = \inf_{m \geq 1} |t - S_m|$ from t to the nearest arrival time. Since $S_m \to \infty$, the infimum is achieved. Further, $D_t = 0$ if and only if $S_m = t$ for some m, and since by hypothesis there are no fixed discontinuities, the probability of this is 0: $P[D_t = 0] = 0$. Choose δ_t so that $0 < \delta_t < n^{-1}$ and $P[D_t \leq \delta_t] < n^{-1}$. The intervals $(t - \delta_t, t + \delta_t)$ for $t' \leq t \leq t''$ cover $[t', t'']$. Choose a finite subcover, and in (23.18) take the t_{nk} for $0 < k < r_n$ to be the endpoints (of intervals in the subcover) that are contained in (t', t''). By the construction,

$$\max_{1 \leq k \leq r_n} \left(t_{nk} - t_{n,k-1}\right) \to 0, \tag{23.22}$$

and the probability that $(t_{n,k-1}, t_{nk}]$ contains some S_m is less than n^{-1}. This gives a sequence of partitions satisfying (23.20). Inserting more points in a partition cannot increase the maxima in (23.20) and (23.22), and so it can be arranged that each partition refines the preceding one.

To prove (23.21) it is enough (Theorem 4.1) to show that the limit superior of the sets involved has probability 0. It is in fact empty: If for infinitely many n, $N_{t_{nk}}(\omega) - N_{t_{n,k-1}}(\omega) \geq 2$ holds for some $k \leq r_n$, then by (23.22), $N_t(\omega)$ as a

function of t has in $[t', t'']$ discontinuity points (arrival times) arbitrarily close together, which requires $S_m(\omega) \in [t', t'']$ for infinitely many m, in violation of Condition 0°.

It remains to prove (23.19). If Z_{nk} and Z_n are defined as above and $p_{nk} = P[Z_{nk} = 1]$, then the sum in (23.19) is $\sum_k p_{nk} = E[Z_n]$. Since $Z_{n+1} \geq Z_n$, $\sum_k p_{nk}$ is nondecreasing in n. Now

$$P[N_{t''} - N_{t'} = 0] = P[Z_{nk} = 0,\ k \leq r_n]$$
$$= \prod_{k=1}^{r_n} (1 - p_{nk}) \leq \exp\left[- \sum_{k=1}^{r_n} p_{nk} \right].$$

If the left-hand side here is positive, this puts an upper bound on $\sum_k p_{nk}$, and (23.19) follows. But suppose $P[N_{t''} - N_{t'} = 0] = 0$. If s is the midpoint of t' and t'', then since the increments are independent, one of $P[N_s - N_{t'} = 0]$ and $P[N_{t''} - N_s = 0]$ must vanish. It is therefore possible to find a nested sequence of intervals $[u_m, v_n]$ such that $v_m - u_m \to 0$ and the event $A_m = [N_{v_m} - N_{u_m} \geq 1]$ has probability 1. But then $P(\bigcap_m A_m) = 1$, and if t is the point common to the $[u_m, v_m]$, there is an arrival at t with probability 1, contrary to the assumption that there are no fixed discontinuities. ■

Theorem 23.3 in some cases makes the Poisson model quite plausible. The increments will be essentially independent if the arrivals to time s cannot seriously deplete the population of potential arrivals, so that N_s has for $t > s$ negligible effect on $N_t - N_s$. And the condition that there are no fixed discontinuities is entirely natural. These conditions hold for arrivals of calls at a telephone exchange if the rate of calls is small in comparison with the population of subscribers and calls are not placed at fixed, predetermined times. If the arrival rate is essentially constant, this leads to the following condition.

Condition 3°. (i) *For* $0 < t_1 < \cdots < t_k$ *the increments* $N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}}$ *are independent.*

(ii) *The distribution of* $N_t - N_s$ *depends only on the difference* $t - s$.

Theorem 23.4. *Conditions* 1°, 2°, *and* 3° *are equivalent in the presence of Condition* 0°.

PROOF. Obviously Condition 2° implies 3°. Suppose that Condition 3° holds. If J_t is the saltus at t ($J_t = N_t - \sup_{s<t} N_s$), then $[N_t - N_{t-n^{-1}} \geq 1] \downarrow [J_t \geq 1]$, and it follows by (ii) of Condition 3° that $P[J_t \geq 1]$ is the same for all t. But if the value common to $P[J_t \geq 1]$ is positive, then by the independence of the increments and the second Borel–Cantelli lemma there is probability 1 that $J_t \geq 1$ for infinitely many rational t in $(0, 1)$, for example, which contradicts Condition 0°.

By Theorem 23.3, then, the increments have Poisson distributions. If $f(t)$ is the mean of N_t, then $N_t - N_s$ for $s < t$ must have mean $f(t) - f(s)$ and must by (ii) have mean $f(t-s)$; thus $f(t) = f(s) + f(t-s)$. Therefore, f satisfies Cauchy's functional equation [A20] and, being nondecreasing, must have the form $f(t) = \alpha t$ for $\alpha \geq 0$. Condition 0° makes $\alpha = 0$ impossible. ■

One standard way of deriving the Poisson process is by differential equations.

Condition 4°. If $0 < t_1 < \cdots < t_k$ and if $n_1, \ldots, n_k$ are nonnegative integers, then

$$P\left[N_{t_k+h} - N_{t_k} = 1 \mid N_{t_j} = n_j,\ j \leq k\right] = \alpha h + o(h) \tag{23.23}$$

and

$$P\left[N_{t_k+h} - N_{t_k} \geq 2 \mid N_{t_j} = n_j,\ j \leq k\right] = o(h) \tag{23.24}$$

as $h \downarrow 0$. Moreover, $[N_t\colon t \geq 0]$ has no fixed discontinuities.

The occurrences of $o(h)$ in (23.23) and (23.24) denote functions, say $\phi_1(h)$, and $\phi_2(h)$, such that $h^{-1}\phi_i(h) \to 0$ as $h \downarrow 0$; the ϕ_i may depend a priori on $k, t_1, \ldots, t_k$, and $n_1, \ldots, n_k$ as well as on h. It is assumed in (23.23) and (23.24) that the conditioning events have positive probability, so that the conditional probabilities are well defined.

Theorem 23.5. *Conditions 1° through 4° are all equivalent in the presence of Condition 0°.*

PROOF OF 2° → 4°. For a Poisson process with rate α, the left-hand sides of (23.23) and (23.24) are $e^{-\alpha h}\alpha h$ and $1 - e^{-\alpha h} - e^{-\alpha h}\alpha h$, and these are $\alpha h + o(h)$ and $o(h)$, respectively, because $e^{-\alpha h} = 1 - \alpha h + o(h)$. And by the argument in the preceding proof, the process has no fixed discontinuities. ■

PROOF OF 4° → 2°. Fix k, the t_j, and the n_j; denote by A the event $[N_{t_j} = n_j,\ j \leq k]$; and for $t \geq 0$ put $p_n(t) = P[N_{t_k+t} - N_{t_k} = n \mid A]$. It will be shown that

$$p_n(t) = e^{-\alpha t}\frac{(\alpha t)^n}{n!}, \qquad n = 0, 1, \ldots. \tag{23.25}$$

This will also be proved for the case in which $p_n(t) = P[N_t = n]$. Condition 2° will then follow by induction.

If $t > 0$ and $|t - s| < n^{-1}$, then

$$|P[N_t = n] - P[N_s = n]| \le P[N_t \ne N_s] \le P[N_{t+n^{-1}} - N_{t-n^{-1}} \ge 1].$$

As $n \to \infty$, the right side here decreases to the probability of a discontinuity at t, which is 0 by hypothesis. Thus $P[N_t = n]$ is continuous at t. The same kind of argument works for conditional probabilities and for $t = 0$, and so $p_n(t)$ is continuous for $t \ge 0$.

To simplify the notation, put $D_t = N_{t_k+t} - N_{t_k}$. If $D_{t+h} = n$, then $D_t = m$ for some $m \le n$. If $t > 0$, then by the rules for conditional probabilities,

$$\begin{aligned} p_n(t+h) = {} & p_n(t) P[D_{t+h} - D_t = 0 | A \cap [D_t = n]] \\ & + p_{n-1}(t) P[D_{t+h} - D_t = 1 | A \cap [D_t = n-1]] \\ & + \sum_{m=0}^{n-2} p_m(t) P[D_{t+h} - D_t = n - m | A \cap [D_t = m]]. \end{aligned}$$

For $n \le 1$, the final sum is absent, and for $n = 0$, the middle term is absent as well. This holds in the case $p_n(t) = P[N_t = n]$ if $D_t = N_t$ and $A = \Omega$. (If $t = 0$, some of the conditioning events here are empty; hence the assumption $t > 0$.) By (23.24), the final sum is $o(h)$ for each fixed n. Applying (23.23) and (23.24) now leads to

$$p_n(t+h) = p_n(t)(1 - \alpha h) + p_{n-1}(t)\alpha h + o(h),$$

and letting $h \downarrow 0$ gives

$$p_n'(t) = -\alpha p_n(t) + \alpha p_{n-1}(t). \tag{23.26}$$

In the case $n = 0$, take $p_{-1}(t)$ to be identically 0. In (23.26), $t > 0$ and $p_n'(t)$ is a right-hand derivative. But since $p_n(t)$ and the right side of the equation are continuous on $[0, \infty)$, (23.26) holds also for $t = 0$ and $p_n'(t)$ can be taken as a two-sided derivative for $t > 0$ [A22].

Now (23.26) gives [A23]

$$p_n(t) = e^{-\alpha t}\left[p_n(0) + \alpha \int_0^t p_{n-1}(s) e^{\alpha s}\, ds \right].$$

Since $p_n(0)$ is 1 or 0 as $n = 0$ or $n > 0$, (23.25) follows by induction on n. ■

Stochastic Processes

The Poisson process $[N_t\colon t \geq 0]$ is one example of a *stochastic process*—that is, a collection of random variables (on some probability space $(\Omega, \mathscr{F}, P)$) indexed by a parameter regarded as representing time. In the Poisson case, time is *continuous*. In some cases the time is *discrete*: Section 7 concerns the sequence $\{F_n\}$ of a gambler's fortunes; there n represents time, but time that increases in jumps.

Part of the structure of a stochastic process is specified by its *finite-dimensional distributions*. For any finite sequence $t_1, \ldots, t_k$ of time points, the k-dimensional random vector $(N_{t_1}, \ldots, N_{t_k})$ has a distribution $\mu_{t_1 \ldots t_k}$ over R^k. These measures $\mu_{t_1 \ldots t_k}$ are the finite-dimensional distributions of the process. Condition 2° of this section in effect specifies them for the Poisson case:

$$P\left[N_{t_j} = n_j,\ j \leq k\right] = \prod_{j=1}^{k} e^{-\alpha(t_j - t_{j-1})} \frac{\left(\alpha(t_j - t_{j-1})\right)^{n_j - n_{j-1}}}{(n_j - n_{j-1})!} \tag{23.27}$$

if $0 \leq n_1 \leq \cdots \leq n_k$ and $0 \leq t_1 < \cdots < t_k$ (take $n_0 = t_0 = 0$).

The finite-dimensional distributions do not, however, contain all the mathematically interesting information about the process in the case of continuous time. Because of (23.3), (23.4), and the definition (23.5), for each fixed ω, $N_t(\omega)$ as a function of t has the regularity properties given in the second version of Condition 0°. These properties are used in an essential way in the proofs.

Suppose that $f(t)$ is t or 0 according as t is rational or irrational. Let N_t be defined as before, and let

$$M_t(\omega) = N_t(\omega) + f\left(t + X_1(\omega)\right). \tag{23.28}$$

If R is the set of rationals, then $P[\omega\colon f(t + X_1(\omega)) \neq 0] = P[\omega\colon X_1(\omega) \in R - t] = 0$ for each t because $R - t$ is countable and X_1 has a density. Thus $P[M_t = N_t] = 1$ for each t, and so the stochastic process $[M_t\colon t \geq 0]$ has the same finite-dimensional distributions as $[N_t\colon t \geq 0]$. For ω fixed, however, $M_t(\omega)$ as a function of t is everywhere discontinuous and is neither monotone nor exclusively integer-valued.

The functions obtained by fixing ω and letting t vary are called the *path functions* or *sample paths* of the process. The example above shows that the finite-dimensional distributions do not suffice to determine the character of the path functions. In specifying a stochastic process as a model for some phenomenon, it is natural to place conditions on the character of the sample paths as well as on the finite-dimensional distributions. Condition 0° was imposed throughout this section to ensure that the sample paths are nondecreasing, right-continuous, integer-valued step functions, a natural condition if N_t is to represent the number of events in $[0, t]$. Stochastic processes in continuous time are studied further in Chapter 7.

PROBLEMS

Assume the Poisson processes here satisfy Condition 0° as well as Condition 1°.

23.1. Show that the minimum of independent exponential waiting times is again exponential and that the parameters add.

23.2. 20.17↑ Show that the time S_n of the nth event in a Poisson stream has the gamma density $f(x;\alpha,n)$ as defined by (20.47). This is sometimes called the *Erlang* density.

23.3. Let $A_t = t - S_{N_t}$ be the time back to the most recent event in the Poisson stream (or to 0), and let $B_t = S_{N_t+1} - t$ be the time forward to the next event. Show that A_t and B_t are independent, that B_t is distributed as X_1 (exponentially with parameter α), and that A_t is distributed as $\min\{X_1, t\}$: $P[A_t \le t]$ is 0, $1 - e^{-\alpha x}$, or 1 as $x < 0$, $0 \le x < t$, or $x \ge t$.

23.4. ↑ Let $L_t = A_t + B_t = S_{N_t+1} - S_{N_t}$ be the length of the interarrival interval covering t.

(a) Show that L_t has density

$$d_t(x) = \begin{cases} \alpha^2 x e^{-\alpha x} & \text{if } 0 < x < t, \\ \alpha(1+\alpha t)e^{-\alpha x} & \text{if } x \ge t. \end{cases}$$

(b) Show that $E[L_t]$ converges to $2E[X_1]$ as $t \to \infty$. This seems paradoxical because L_t is one of the X_n. Give an intuitive resolution of the apparent paradox.

23.5. *Merging Poisson streams.* Define a process $\{N_t\}$ by (23.5) for a sequence $\{X_n\}$ of random variables satisfying (23.4). Let $\{X_n'\}$ be a second sequence of random variables, on the same probability space, satisfying (23.4), and define $\{N_t'\}$ by $N_t' = \max[n\colon X_1' + \cdots + X_n' \le t]$. Define $\{N_t''\}$ by $N_t'' = N_t + N_t'$. Show that, if $\sigma(X_1, X_2, \ldots)$ and $\sigma(X_1', X_2', \ldots)$ are independent and $\{N_t\}$ and $\{N_t'\}$ are Poisson processes with respective rates α and β, then $\{N_t''\}$ is a Poisson process with rate $\alpha + \beta$.

23.6. ↑ The nth and $(n+1)$st events in the process $\{N_t\}$ occur at times S_n and S_{n+1}.

(a) Find the distribution of the number $N'_{S_{n+1}} - N'_{S_n}$ of events in the other process during this time interval.

(b) Generalize to $N'_{S_m} - N'_{S_n}$.

23.7. Suppose that $X_1, X_2, \ldots$ are independent and exponentially distributed with parameter α, so that (23.5) defines a Poisson process $\{N_t\}$. Suppose that $Y_1, Y_2, \ldots$ are independent and identically distributed and that $\sigma(X_1, X_2, \ldots)$ and $\sigma(Y_1, Y_2, \ldots)$ are independent. Put $Z_t = \sum_{k \le N_t} Y_k$. This is the *compound Poisson process*. If, for example, the event at time S_n in the original process

represents an insurance claim, and if Y_n represents the amount of the claim, then Z_t represents the total claims to time t.

(a) If $Y_k = 1$ with probability 1, then $\{Z_t\}$ is an ordinary Poisson process.

(b) Show that $\{Z_t\}$ has independent increments and that $Z_{s+t} - Z_s$ has the same distribution as Z_t.

(c) Show that, if Y_k assumes the values 1 and 0 with probabilities p and $1-p$ $(0 < p < 1)$, then $\{Z_t\}$ is a Poisson process with rate $p\alpha$.

23.8. Suppose a process satisfies Condition 0° and has independent, Poisson-distributed increments and no fixed discontinuities. Show that it has the form $\{N_{\varphi(t)}\}$, where $\{N_t\}$ is a standard Poisson process and φ is a nondecreasing, continuous function on $[0, \infty)$ with $\varphi(0) = 0$.

23.9. If the waiting times X_n are independent and exponentially distributed with parameter α, then $S_n/n \to \alpha^{-1}$ with probability 1, by the strong law of large numbers. From $\lim_{t\to\infty} N_t = \infty$ and $S_{N_t} \le t < S_{N_t+1}$ deduce that $\lim_{t\to\infty} N_t/t = \alpha$ with probability 1.

23.10. ↑ (a) Suppose that $X_1, X_2, \ldots$ are positive, and assume directly that $S_n/n \to m$ with probability 1, as happens if the X_n are independent and identically distributed with mean m. Show that $\lim_t N_t/t = 1/m$ with probability 1.

(b) Suppose now that $S_n/n \to \infty$ with probability 1, as happens if the X_n are independent and identically distributed and have infinite mean. Show that $\lim_t N_t/t = 0$ with probability 1.

The results in Problem 23.10 are theorems in *renewal theory*: A component of some mechanism is replaced each time it fails or wears out. The X_n are the lifetimes of the successive components, and N_t is the number of replacements, or renewals, to time t.

23.11. 20.7 23.10↑ Consider a persistent, irreducible Markov chain, and for a fixed state j let N_n be the number of passages through j up to time n. Show that $N_n/n \to 1/m$ with probability 1, where $m = \sum_{k=1}^{\infty} k f_{jj}^{(k)}$ is the mean return time (replace $1/m$ by 0 if this mean is infinite). See Lemma 3 in Section 8.

23.12. Suppose that X and Y have Poisson distributions with parameters α and β. Show that $|P[X = i] - P[Y = i]| \le |\alpha - \beta|$. *Hint:* Suppose that $\alpha < \beta$, and represent Y as $X + D$, where X and D are independent and have Poisson distributions with parameters α and $\beta - \alpha$.

23.13. ↑ Use the methods in the proof of Theorem 23.2 to show that the error in (23.15) is bounded uniformly in i by $|\lambda - \lambda_n| + \lambda_n \max_k p_{nk}$.

SECTION 24. THE ERGODIC THEOREM*

Even though chance necessarily involves the notion of change, the laws governing the change may themselves remain constant as time passes: If time

*This section may be omitted. There is more on ergodic theory in Section 36.

does not alter the roulette wheel, the gambler's fortunes fluctuate according to constant probability laws. The ergodic theorem is a version of the strong law of large numbers general enough to apply to any system governed by probability laws that are invariant in time.

Measure-Preserving Transformations

Let $(\Omega, \mathscr{F}, P)$ be a probability space. A mapping $T: \Omega \to \Omega$ is a *measure-preserving transformation* if it is measurable $\mathscr{F}/\mathscr{F}$ and $P(T^{-1}A) = P(A)$ for all A in $\mathscr{F}$, from which it follows that $P(T^{-n}A) = P(A)$ for $n \geq 0$. If, further, T is a one-to-one mapping onto Ω and the point mapping T^{-1} is measurable $\mathscr{F}/\mathscr{F}$ ($TA \in \mathscr{F}$ for $A \in \mathscr{F}$), then T is *invertible*; in this case T^{-1} automatically preserves measure: $P(A) = P(T^{-1}TA) = P(TA)$.

The first result is a simple consequence of the π–λ theorem (or Theorems 13.1(i) and 3.3).

Lemma 1. *If $\mathscr{P}$ is a π-system generating $\mathscr{F}$, and if $T^{-1}A \in \mathscr{F}$ and $P(T^{-1}A) = P(A)$ for $A \in \mathscr{P}$, then T is a measure-preserving transformation.*

Example 24.1. *The Bernoulli shift.* Let S be a finite set, and consider the space S^∞ of sequences (2.15) of elements of S. Define the *shift* T by

$$T\omega = (z_2(\omega), z_3(\omega), \dots); \tag{24.1}$$

the first element of $\omega = (z_1(\omega), z_2(\omega), \dots)$ is lost, and T shifts the remaining elements one place to the left: $z_k(T\omega) = z_{k+1}(\omega)$ for $k \geq 1$. If A is a cylinder (2.17), then

$$\begin{aligned} T^{-1}A &= \left[\omega: (z_2(\omega), \dots, z_{n+1}(\omega)) \in H\right] \\ &= \left[\omega: (z_1(\omega), \dots, z_{n+1}(\omega)) \in S \times H\right] \end{aligned} \tag{24.2}$$

is another cylinder, and since the cylinders generate the basic σ-field $\mathscr{C}$, T is measurable $\mathscr{C}/\mathscr{C}$.

For probabilities p_u on S (nonnegative and summing to 1), define product measure P on the field $\mathscr{C}_0$ of cylinders by (2.21). Then P is consistently defined and countably additive (Theorem 2.3) and hence extends to a probability measure on $\mathscr{C} = \sigma(\mathscr{C}_0)$. Since the thin cylinders (2.16) form a π-system that generates $\mathscr{C}$, P is completely determined by the probabilities it assigns to them:

$$P\left[\omega: (z_1(\omega), \dots, z_n(\omega)) = (u_1, \dots, u_n)\right] = p_{u_1} \cdots p_{u_n}. \tag{24.3}$$

If A is the thin cylinder on the left here, then by (24.2),

$$(24.4) \qquad P(T^{-1}A) = \sum_{u \in S} p_u p_{u_1} \cdots p_{u_n} = p_{u_1} \cdots p_{u_n} = P(A),$$

and it follows by the lemma that T preserves P. This T is the *Bernoulli shift*.

If $A = [\omega\colon z_1(\omega) = u]$ $(u \in S)$, then $I_A(T^{k-1}\omega)$ is 1 or 0 according as $z_k(\omega)$ is u or not. Since by (24.3) the events $[\omega\colon z_k(\omega) = u]$ are independent, each with probability p_u, the random variables $I_A(T^{k-1}\omega)$ are independent and take the value 1 with probability $p_u = P(A)$. By the strong law of large numbers, therefore,

$$(24.5) \qquad \lim_n \frac{1}{n} \sum_{k=1}^{n} I_A(T^{k-1}\omega) = P(A)$$

with probability 1. ■

Example 24.1 gives a model for independent trials, and that T preserves P means the probability laws governing the trials are invariant in time. In the present section, it is this invariance of the probability laws that plays the fundamental role; independence is a side issue.

The *orbit* under T of the point ω is the sequence $(\omega, T\omega, T^2\omega, \ldots)$, and (24.5) can be expressed by saying that the orbit enters the set A with asymptotic relative frequency $P(A)$. For $A = [\omega\colon (z_1(\omega), z_2(\omega)) = (u_1, u_2)]$, the $I_A(T^{k-1}\omega)$ are not independent, but (24.5) holds anyway. In fact, for the Bernoulli shift, (24.5) holds with probability 1 *whatever* A may be $(A \in \mathcal{F})$. This is one of the consequences of the ergodic theorem (Theorem 24.1). What is more, according to this theorem the limit in (24.5) exists with probability 1 (although it may not be constant in ω) if T is an arbitrary measure-preserving transformation on an arbitrary probability space, of which there are many examples.

Example 24.2. *The Markov shift.* Let $P = [p_{ij}]$ be a stochastic matrix with rows and columns indexed by the finite set S, and let π_i be probabilities on S. Replace (2.21) by $P(A) = \sum_H \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n}$. The argument in Section 2 showing that product measure is consistently defined and finitely additive carries over to this new measure: since the rows of the transition matrix add to 1,

$$\sum_{u_{n+1} \cdots u_m} p_{u_n u_{n+1}} \cdots p_{u_{m-1} u_m} = 1,$$

and so the argument involving (2.23) goes through. The new measure is again

countably additive on $\mathscr{C}_0$ (Theorem 2.3) and so extends to $\mathscr{C}$. This probability measure P on $\mathscr{C}$ is uniquely determined by the condition

$$(24.6)\quad P[\omega: (z_1(\omega), \ldots, z_n(\omega)) = (u_1, \ldots, u_n)] = \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n}.$$

Thus the coordinate functions $z_n(\cdot)$ are a Markov chain with transition probabilities p_{ij} and initial probabilities π_i.

Suppose that the π_i (until now unspecified) are stationary: $\sum_i \pi_i p_{ij} = \pi_j$. Then

$$\sum_{u \in S} \pi_u p_{uu_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n} = \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n},$$

and it follows (see (24.4)) that T preserves P. Under the measure P specified by (24.6), T is the *Markov shift*. ■

The shift T, qua point transformation on S^∞, is the same in Examples 24.1 and 24.2. A measure-preserving transformation, however, is the point transformation together with the σ-field with respect to which it is measurable and the measure it preserves.

Example 24.3. Let P be Lebesgue measure λ on the unit interval, and take $T\omega = 2\omega \pmod 1$:

$$T\omega = \begin{cases} 2\omega & \text{if } 0 < \omega \le \frac{1}{2}, \\ 2\omega - 1 & \text{if } \frac{1}{2} < \omega \le 1. \end{cases}$$

If ω has nonterminating dyadic expansion $\omega = .d_1(\omega)d_2(\omega)\ldots$, then $T\omega = .d_2(\omega)d_3(\omega)\ldots$: T shifts the digits one place to the left—compare (24.1). Since $T^{-1}(0, x] = (0, \frac{1}{2}x] \cup (\frac{1}{2}, \frac{1}{2} + \frac{1}{2}x]$, it follows by Lemma 1 that T preserves Lebesgue measure. This is the *dyadic* transformation. ■

Example 24.4. Let Ω be the unit circle in the complex plane, let $\mathscr{F}$ be the σ-field generated by the arcs, and let P be normalized circular Lebesgue measure: map $[0, 1)$ to the unit circle by $\phi(x) = e^{2\pi i x}$ and define P by $P(A) = \lambda(\phi^{-1}A)$. For a fixed c in Ω, let $T\omega = c\omega$. Since T is effectively the *rotation* of the circle through the angle $\arg c$, T preserves P. The rotation turns out to have radically different properties according as c is a root of unity or not. ■

Ergodicity

The $\mathscr{F}$-set A is *invariant* under T if $T^{-1}A = A$; it is a *nontrivial* invariant set if $0 < P(A) < 1$. And T is by definition *ergodic* if there are in $\mathscr{F}$ no

nontrivial invariant sets. A measurable function f is invariant if $f(T\omega) = f(\omega)$ for all ω; A is invariant if and only if I_A is.

The *ergodic theorem*:

Theorem 24.1. *Suppose that T is a measure-preserving transformation on $(\Omega, \mathcal{F}, P)$ and that f is measurable and integrable. Then*

$$\lim_n \frac{1}{n} \sum_{k=1}^{n} f(T^{k-1}\omega) = \hat{f}(\omega) \tag{24.7}$$

with probability 1, *where $\hat{f}$ is invariant and integrable and $E[\hat{f}] = E[f]$. If T is ergodic, then $\hat{f} = E[f]$ with probability* 1.

This will be proved later in the section. In Section 34, $\hat{f}$ will be identified as a conditional expected value (see Example 34.3).

If $f = I_A$, (24.7) becomes

$$\lim_n \frac{1}{n} \sum_{k=1}^{n} I_A(T^{k-1}\omega) = \hat{I}_A(\omega), \tag{24.8}$$

and in the ergodic case,

$$\hat{I}_A(\omega) = P(A) \tag{24.9}$$

with probability 1. If A is invariant, then $\hat{I}_A(\omega)$ is 1 on A and 0 on A^c, and so the limit can certainly be nonconstant if T is not ergodic.

Example 24.5. Take $\Omega = \{a, b, c, d, e\}$ and $\mathcal{F} = 2^{\Omega}$. If T is the cyclic permutation $T = (a, b, c, d, e)$ and T preserves P, then P assigns equal probabilities to the five points. Since $\varnothing$ and Ω are the only invariant sets, T is ergodic. It is easy to check (24.8) and (24.9) directly.

The transformation $T = (a, b, c)(d, e)$, a product of two cycles, preserves P if and only if a, b, c have equal probabilities and d, e have equal probabilities. If the probabilities are all positive, then since $\{a, b, c\}$ is invariant, T is not ergodic. If, say, $A = \{a, d\}$, the limit in (24.8) is $\frac{1}{3}$ on $\{a, b, c\}$ and $\frac{1}{2}$ on $\{d, e\}$. This illustrates the essential role of ergodicity. ■

The coordinate functions $z_n(\cdot)$ in Example 24.1 are independent, and hence by Kolmogorov's zero-one law every set in the tail field $\mathcal{T} = \bigcap_n \sigma(z_n, z_{n+1}, \ldots)$ has probability 0 or 1. (That the z_n take values in the abstract set S does not affect the arguments.) If $A \in \sigma(z_1, \ldots, z_k)$, then $T^{-n}A \in \sigma(z_{n+1}, \ldots, z_{n+k}) \subset \sigma(z_{n+1}, z_{n+2}, \ldots)$; since this is true for each k, $A \in \mathcal{F} = \sigma(z_1, z_2, \ldots)$ implies (Theorem 13.1(i)) $T^{-n}A \in \sigma(z_{n+1}, z_{n+2}, \ldots)$. For A invariant, it follows that $A \in \mathcal{T}$: The Bernoulli shift is ergodic. Thus

the ergodic theorem does imply that (24.5) holds with probability 1, whatever A may be.

The ergodicity of the Bernoulli shift can be proved in a different way. If $A = [(z_1, \ldots, z_n) = u]$ and $B = [(z_1, \ldots, z_k) = v]$ for an n-tuple u and a k-tuple v, and if P is given by (24.3), then $P(A \cap T^{-n}B) = P(A)P(B)$ because $T^{-n}B = [(z_{n+1}, \ldots, z_{n+k}) = v]$. Fix n and A, and use the π–λ theorem to show that this holds for all B in $\mathscr{F}$. If B is invariant, then $P(A \cap B) = P(A)P(B)$ holds for the A above, and hence (π–λ again) holds for all A. Taking $B = A$ shows that $P(B) = (P(B))^2$ for invariant B, and $P(B)$ is 0 or 1. This argument is very close to the proof of the zero–one law, but a modification of it gives a criterion for ergodicity that applies to the Markov shift and other transformations.

Lemma 2. *Suppose that $\mathscr{P} \subset \mathscr{F}_0 \subset \mathscr{F}$, where $\mathscr{F}_0$ is a field, every set in $\mathscr{F}_0$ is a finite or countable disjoint union of $\mathscr{P}$-sets, and $\mathscr{F}_0$ generates $\mathscr{F}$. Suppose further that there exists a positive c with this property: For each A in $\mathscr{P}$ there is an integer n_A such that*

$$P(A \cap T^{-n_A}B) \geq cP(A)P(B) \tag{24.10}$$

for all B in $\mathscr{P}$. Then $T^{-1}C = C$ implies that $P(C)$ is 0 or 1.

It is convenient not to require that T preserve P; but if it does, then it is an ergodic measure-preserving transformation. In the argument just given, $\mathscr{P}$ consists of the thin cylinders, $n_A = n$ if $A = [(z_1, \ldots, z_n) = u]$, $c = 1$, and $\mathscr{F}_0 = \mathscr{C}_0$ is the class of cylinders.

PROOF. Since every $\mathscr{F}_0$-set is a disjoint union of $\mathscr{P}$-sets, (24.10) holds for $B \in \mathscr{F}_0$ (and $A \in \mathscr{P}$). Since for fixed A the class of B satisfying (24.10) is monotone, it contains $\mathscr{F}$ (Theorem 3.4). If B is invariant, it follows that $P(A \cap B) \geq cP(A)P(B)$ for A in $\mathscr{P}$. But then, by the same argument, the inequality holds for all A in $\mathscr{F}$. Take $A = B^c$: If B is invariant, then $P(B^c)P(B) = 0$ and hence $P(B)$ is 0 or 1. ■

To treat the Markov shift, take $\mathscr{F}_0$ to consist of the cylinders and $\mathscr{P}$ to consist of the thin ones. If $A = [(z_1, \ldots, z_n) = (u_1, \ldots, u_n)]$, $n_A = n + m - 1$, and $B = [(z_1, \ldots, z_k) = (v_1, \ldots, v_k)]$, then

$$\begin{aligned} P(A)P(B) &= \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n} \pi_{v_1} p_{v_1 v_2} \cdots p_{v_{k-1} v_k}, \\ P(A \cap T^{-n_A}B) &= \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n} p^{(m)}_{u_n v_1} p_{v_1 v_2} \cdots p_{v_{k-1} v_k}. \end{aligned} \tag{24.11}$$

The lemma will apply if there exist an integer m and a positive c such that $p^{(m)}_{ij} \geq c\pi_j$ for all i and j. By Theorem 8.9 (or Lemma 2, p. 125), there is in the irreducible, aperiodic case an m such that all $p^{(m)}_{ij}$ are positive; take c

less than the minimum. By Lemma 2, the corresponding Markov shift is ergodic.

Example 24.6. Maps preserve ergodicity. Suppose that $\psi\colon \Omega \to \Omega$ is measurable $\mathscr{F}/\mathscr{F}$ and commutes with T in the sense that $\psi T\omega = T\psi\omega$. If T preserves P, it also preserves $P\psi^{-1}$: $P\psi^{-1}(T^{-1}A) = P(T^{-1}\psi^{-1}A) = P\psi^{-1}(A)$. And if T is ergodic under P, it is also ergodic under $P\psi^{-1}$: if A is invariant, so is $\psi^{-1}A$, and hence $P\psi^{-1}(A)$ is 0 or 1. These simple observations are useful in studying the ergodicity of stochastic processes (Theorem 36.4). ■

Ergodicity of Rotations

The dyadic transformation, Example 24.3, is essentially the same as the Bernoulli shift. In any case, it is easy to use the zero–one law or Lemma 2 to show that it is ergodic. From this and the ergodic theorem, the normal number theorem follows once again.

Consider the rotations of Example 24.4. If the complex number c defining the rotation ($T\omega = c\omega$) is -1, then the set consisting of the first and third quadrants is a nontrivial invariant set, and hence T is not ergodic. A similar construction shows that T is nonergodic whenever c is a root of unity.

In the opposite case, c *is* ergodic. In the first place, it is an old number-theoretic fact due to Kronecker that if c is not a root of unity then the orbit $(\omega, c\omega, c^2\omega, \dots)$ of every ω is dense. Since the orbits are rotations of one another, it suffices to prove that the orbit $(1, c, c^2, \dots)$ of 1 is dense. But if c is not a root of unity, then the elements of this orbit are all distinct and hence by compactness have a limit point ω_0. For arbitrary ϵ, there are distinct points c^n and c^{n+k} within $\epsilon/2$ of ω_0 and hence within ϵ of each other (distance measured along the arc). But then, since the distance from c^{n+jk} to $c^{n+(j+1)k}$ is the same as that from c^n to c^{n+k}, it is clear that for some m the points $c^n, c^{n+k}, \dots, c^{n+mk}$ form a chain which extends all the way around the circle and in which the distance from one point to the next is less than ϵ. Thus every point on the circle is within ϵ of some point of the orbit $(1, c, c^2, \dots)$, which is indeed dense.

To use this result to prove ergodicity, suppose that A is invariant and $P(A) > 0$. To show that $P(A)$ must then be 1, observe first that for arbitrary ϵ there is an arc I, of length at most ϵ, satisfying $P(A \cap I) > (1-\epsilon)P(A)$. Indeed, A can be covered by a sequence $I_1, I_2, \dots$ of arcs for which $P(A)/(1-\epsilon) > \sum_n P(I_n)$; the arcs can be taken disjoint and of length less than ϵ. Since $\sum_n P(A \cap I_n) = P(A) > (1-\epsilon)\sum_n P(I_n)$, there is an n for which $P(A \cap I_n) > (1-\epsilon)P(I_n)$: take $I = I_n$. Let I have length α; $\alpha \le \epsilon$.

Since A is invariant and T is invertible and preserves P, it follows that $P(A \cap T^n I) \ge (1-\epsilon)P(T^n I)$. Suppose the arc I runs from a to b. Let n_1 be arbitrary and, using the fact that $\{T^n a\}$ is dense, choose n_2 so that $T^{n_1}I$ and $T^{n_2}I$ are disjoint and the distance from $T^{n_1}b$ to $T^{n_2}a$ is less than $\epsilon\alpha$. Then choose n_3 so that $T^{n_1}I, T^{n_2}I, T^{n_3}I$ are disjoint and the distance from $T^{n_2}b$ to $T^{n_3}a$ is less than $\epsilon\alpha$. Continue until $T^{n_k}b$ is within α of $T^{n_1}a$ and a further step is impossible. Since the $T^{n_i}I$ are disjoint, $k\alpha \le 1$; and by the construction, the $T^{n_i}I$ cover the circle to within a set of measure $k\epsilon\alpha + \alpha$, which is at most 2ϵ. And now by disjointness,

$$P(A) \ge \sum_{i=1}^{k} P(A \cap T^{n_i}I) \ge (1-\epsilon)\sum_{i=1}^{k} P(T^{n_i}I) \ge (1-\epsilon)(1-2\epsilon).$$

Since ϵ was arbitrary, $P(A)$ must be 1: T is ergodic if c is not a root of unity.†

†For a simple Fourier-series proof, see Problem 26.30.

Proof of the Ergodic Theorem

The argument depends on a preliminary result the statement and proof of which are most clearly expressed in terms of functional operators. For a real function f on Ω, let Uf be the real function with value $(Uf)(\omega) = f(T\omega)$ at ω. If f is integrable, then by change of variable (Theorem 16.13),

$$(24.12) \qquad E[Uf] = \int_\Omega f(T\omega)\,P(d\omega) = \int_\Omega f(\omega)\,PT^{-1}(d\omega) = E[f].$$

And the operator U is nonnegative in the sense that it carries nonnegative functions to nonnegative functions; hence $f \le g$ (pointwise) implies $Uf \le Ug$.

Make these pointwise definitions: $S_0 f = 0$, $S_n f = f + Uf + \cdots + U^{n-1} f$, $M_n f = \max_{0 \le k \le n} S_k f$, and $M_\infty f = \sup_{n \ge 0} S_n f = \sup_{n \ge 0} M_n f$. The *maximal ergodic theorem*:

Theorem 24.2. *If f is integrable, then*

$$(24.13) \qquad \int_{M_\infty f > 0} f\,dP \ge 0.$$

PROOF. Since $B_n = [M_n f > 0] \uparrow [M_\infty f > 0]$, it is enough, by the dominated convergence theorem, to show that $\int_{B_n} f\,dP \ge 0$. On B_n, $M_n f = \max_{1 \le k \le n} S_k f$. Since the operator U is nonnegative, $S_k f = f + US_{k-1} f \le f + UM_n f$ for $1 \le k \le n$, and therefore $M_n f \le f + UM_n f$ on B_n. This and the fact that the function $UM_n f$ is nonnegative imply

$$\begin{aligned} \int_\Omega M_n f\,dP = \int_{B_n} M_n f\,dP &\le \int_{B_n} (f + UM_n f)\,dP \\ &\le \int_{B_n} f\,dP + \int_\Omega UM_n f\,dP = \int_{B_n} f\,dP + \int_\Omega M_n f\,dP, \end{aligned}$$

where the last equality follows from (24.12). Hence $\int_{B_n} f\,dP \ge 0$. ■

Replace f by fI_A. If A is invariant, then $S_n(fI_A) = (S_n f)I_A$, and $M_\infty(fI_A) = (M_\infty f)I_A$, and therefore (24.13) gives

$$(24.14) \qquad \int_{A \cap [M_\infty f > 0]} f\,dP \ge 0 \qquad \text{if } T^{-1}A = A.$$

Now replace f here by $f - \lambda$, λ a constant. Clearly $[M_\infty(f - \lambda) > 0]$ is the set

where for some $n \geq 1$, $S_n(f-\lambda) > 0$, or $n^{-1}S_n f > \lambda$. Let

$$F_\lambda = \left[\omega: \sup_{n\geq 1} \frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega) > \lambda\right]; \tag{24.15}$$

it follows by (24.14) that $\int_{A\cap F_\lambda}(f-\lambda)\,dP \geq 0$, or

$$\lambda P(A\cap F_\lambda) \leq \int_{A\cap F_\lambda} f\,dP \qquad \text{if } T^{-1}A = A. \tag{24.16}$$

The λ here can have either sign.

PROOF OF THEOREM 24.1. To prove that the averages $a_n(\omega) = n^{-1}\sum_{k=1}^{n} f(T^{k-1}\omega)$ converge, consider the set

$$A_{\alpha,\beta} = \left[\omega: \liminf_n a_n(\omega) < \alpha < \beta < \limsup_n a_n(\omega)\right]$$

for $\alpha < \beta$. Since $A_{\alpha,\beta} = A_{\alpha,\beta}\cap F_\beta$ and $A_{\alpha,\beta}$ is invariant, (24.16) gives $\beta P(A_{\alpha,\beta}) \leq \int_{A_{\alpha,\beta}} f\,dP$. The same result with $-f, -\beta, -\alpha$ in place of f, α, β is $\int_{A_{\alpha,\beta}} f\,dP \leq \alpha P(A_{\alpha,\beta})$. Since $\alpha < \beta$, the two inequalities together lead to $P(A_{\alpha,\beta}) = 0$. Take the union over rational α and β: The averages $a_n(\omega)$ converge, that is,

$$\lim_n a_n(\omega) = \hat{f}(\omega) \tag{24.17}$$

with probability 1, where $\hat{f}$ may take the values $\pm\infty$ at certain values of ω.

Because of (24.12), $E[a_n] = E[f]$; if it is shown that the $a_n(\omega)$ are uniformly integrable, then it will follow (Theorem 16.14) that $\hat{f}$ is integrable and $E[\hat{f}] = E[f]$.

By (24.16), $\lambda P(F_\lambda) \leq E[|f|]$. Combine this with the same inequality for $-f$: If $G_\lambda = [\omega: \sup_n |a_n(\omega)| > \lambda]$, then $\lambda P(G_\lambda) \leq 2E[|f|]$ (trivial if $\lambda \leq 0$). Therefore, for positive α and λ,

$$\begin{aligned}
\int_{[|a_n|>\lambda]} |a_n|\,dP &\leq \frac{1}{n}\sum_{k=1}^{n}\int_{G_\lambda} |f(T^{k-1}\omega)|P(d\omega) \\
&\leq \frac{1}{n}\sum_{k=1}^{n}\left(\int_{|f(T^{k-1}\omega)|>\alpha} |f(T^{k-1}\omega)|P(d\omega) + \alpha P(G_\lambda)\right) \\
&= \int_{|f(\omega)|>\alpha} |f(\omega)|P(d\omega) + \alpha P(G_\lambda) \\
&\leq \int_{|f(\omega)|>\alpha} |f(\omega)|P(d\omega) + 2\frac{\alpha}{\lambda}E[|f|].
\end{aligned}$$

Take $\alpha = \lambda^{1/2}$; since f is integrable, the final expression here goes to 0 as

$\lambda \to \infty$. The $a_n(\omega)$ are therefore uniformly integrable, and $E[\hat{f}] = E[f]$. The uniform integrability also implies $E[|a_n - \hat{f}|] \to 0$.

Set $\hat{f}(\omega) = 0$ outside the set where the $a_n(\omega)$ have a finite limit. Then (24.17) still holds with probability 1, and $\hat{f}(T\omega) = \hat{f}(\omega)$. Since $[\omega\colon \hat{f}(\omega) \le x]$ is invariant, in the ergodic case its measure is either 0 or 1; if x_0 is the infimum of the x for which it is 1, then $\hat{f}(\omega) = x_0$ with probability 1, and from $x_0 = E[\hat{f}] = E[f]$ it follows that $\hat{f}(\omega) = E[f]$ with probability 1. ■

The Continued-Fraction Transformation

Let Ω consist of the irrationals in the unit interval, and for x in Ω let $Tx = \{1/x\}$ and $a_1(x) = [1/x]$ be the fractional and integral parts of $1/x$. This defines a mapping

$$Tx = \left\{\frac{1}{x}\right\} = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor = \frac{1}{x} - a_1(x) \tag{24.18}$$

of Ω into itself, a mapping associated with the continued-fraction expansion of x [A36]. Concentrating on *irrational* x avoids some trivial details connected with the rational case, where the expansion is finite; it is an inessential restriction because the interest here centers on results of the almost-everywhere kind.

For $x \in \Omega$ and $n \ge 1$ let $a_n(x) = a_1(T^{n-1}x)$ be the nth partial quotient, and define integer-valued functions $p_n(x)$ and $q_n(x)$ by the recursions

(24.19)

$$\begin{aligned} &p_{-1}(x) = 1, \quad && p_0(x) = 0, && p_n(x) = a_n(x)p_{n-1}(x) + p_{n-2}(x), && n \ge 1, \\ &q_{-1}(x) = 0, && q_0(x) = 1, && q_n(x) = a_n(x)q_{n-1}(x) + q_{n-2}(x), && n \ge 1. \end{aligned}$$

Simple induction arguments show that

$$p_{n-1}(x)q_n(x) - p_n(x)q_{n-1}(x) = (-1)^n, \qquad n \ge 0, \tag{24.20}$$

and [A37: (27)]

$$x = \underline{1\lfloor a_1(x) \rfloor} + \cdots + \underline{1\lfloor a_{n-1}(x) \rfloor} + \underline{1\lfloor a_n(x) + T^n x \rfloor}, \qquad n \ge 1. \tag{24.21}$$

It also follows inductively [A36: (26)] that

$$\begin{aligned} &\underline{1\lfloor a_1(x) \rfloor} + \cdots + \underline{1\lfloor a_{n-1}(x) \rfloor} + \underline{1\lfloor a_n(x) + t \rfloor} \\ &\qquad = \frac{p_n(x) + tp_{n-1}(x)}{q_n(x) + tq_{n-1}(x)}, \qquad n \ge 1, \quad 0 \le t \le 1. \end{aligned} \tag{24.22}$$

Taking $t = 0$ here gives the formula for the nth convergent:

$$(24.23) \qquad \underline{1}\big/a_1(x) + \cdots + \underline{1}\big/a_n(x) = \frac{p_n(x)}{q_n(x)}, \qquad n \geq 1,$$

where, as follows from (24.20), $p_n(x)$ and $q_n(x)$ are relatively prime. By (24.21) and (24.22),

$$(24.24) \qquad x = \frac{p_n(x) + (T^n x)p_{n-1}(x)}{q_n(x) + (T^n x)q_{n-1}(x)}, \qquad n \geq 0,$$

which, together with (24.20), implies[†]

$$(24.25) \qquad x - \frac{p_n(x)}{q_n(x)} = \frac{(-1)^n}{q_n(x)\big((T^n x)^{-1}q_n(x) + q_{n-1}(x)\big)}, \qquad n \geq 0.$$

Thus the convergents for even n fall to the left of x, and those for odd n fall to the right. And since (24.19) obviously implies that $q_n(x)$ goes to infinity with n, the convergents $p_n(x)/q_n(x)$ do converge to x: Each irrational x in $(0, 1)$ has the infinite simple continued-fraction representation

$$(24.26) \qquad x = \underline{1}\big/a_1(x) + \underline{1}\big/a_2(x) + \cdots .$$

The representation is unique [A36: (35)], and $Tx = \underline{1}\big/a_2(x) + \underline{1}\big/a_3(x) + \cdots$: T shifts the partial quotients in the same way the dyadic transformation (Example 24.3) shifts the digits of the dyadic expansion. Since the continued-fraction transformation turns out to be ergodic, it can be used to study the continued-fraction algorithm.

Suppose now that $a_1, a_2, \ldots$ are positive integers and define p_n and q_n by the recursions (24.19) without the argument x. Then (24.20) again holds (without the x), and so $p_n/q_n - p_{n-1}/q_{n-1} = (-1)^{n+1}/q_{n-1}q_n$, $n \geq 1$. Since q_n increases to infinity, the right side here is the nth term of a convergent alternating series. And since $p_0/q_0 = 0$, the nth partial sum is p_n/q_n, which therefore converges to some limit: Every simple infinite continued fraction converges, and [A36: (36)] the limit is an irrational in $(0, 1)$.

Let $\Delta_{a_1 \ldots a_n}$ be the set of x in Ω such that $a_k(x) = a_k$ for $1 \leq k \leq n$; call it a *fundamental set of rank n*. These sets are analogous to the dyadic intervals and the thin cylinders. For an explicit description of $\Delta_{a_1 \ldots a_n}$—necessary for the proof of ergodicity—consider the function

$$(24.27) \qquad \psi_{a_1 \ldots a_n}(t) = \underline{1}\big/a_1 + \cdots + \underline{1}\big/a_{n-1} + \underline{1}\big/a_n + t.$$

[†]Theorem 1.4 follows from this.

If $x \in \Delta_{a_1 \dots a_n}$, then $x = \psi_{a_1 \dots a_n}(T^n x)$ by (24.21); on the other hand, because of the uniqueness of the partial quotients [A36: (33)], if t is an irrational in the unit interval, then (24.27) lies in $\Delta_{a_1 \dots a_n}$. Thus $\Delta_{a_1 \dots a_n}$ is the image under (24.27) of Ω itself.

Just as (24.22) follows by induction, so does

$$\psi_{a_1 \dots a_n}(t) = \frac{p_n + t p_{n-1}}{q_n + t q_{n-1}}. \tag{24.28}$$

And $\psi_{a_1 \dots a_n}(t)$ is increasing or decreasing in t according as n is even or odd, as is clear from the form of (24.27) (or differentiate in (24.28) and use (24.20)). It follows that

$$\Delta_{a_1 \dots a_n} = \begin{cases} \left(\dfrac{p_n}{q_n}, \dfrac{p_n + p_{n-1}}{q_n + q_{n-1}} \right] \cap \Omega & \text{if } n \text{ is even,} \\ \left(\dfrac{p_n + p_{n-1}}{q_n + q_{n-1}}, \dfrac{p_n}{q_n} \right) \cap \Omega & \text{if } n \text{ is odd.} \end{cases}$$

By (24.20), this set has Lebesgue measure

$$\lambda(\Delta_{a_1 \dots a_n}) = \frac{1}{q_n(q_n + q_{n-1})}. \tag{24.29}$$

The fundamental sets of rank n form a partition of Ω, and their unions form a field $\mathscr{F}_n$; let $\mathscr{F}_0 = \bigcup_{n=1}^{\infty} \mathscr{F}_n$. Then $\mathscr{F}_0$ is the field generated by the class $\mathscr{P}$ of all the fundamental sets, and since each set in $\mathscr{F}_0$ is in some $\mathscr{F}_n$, each is a finite or countable disjoint union of $\mathscr{P}$-sets. Since $q_n \geq 2q_{n-2}$ by (24.19), induction gives $q_n \geq 2^{(n-1)/2}$ for $n \geq 0$. And now (24.29) implies that $\mathscr{F}_0$ generates the σ-field $\mathscr{F}$ of linear Borel sets that are subsets of Ω (use Theorem 10.1(ii)). Thus $\mathscr{P}$, $\mathscr{F}_0$, $\mathscr{F}$ are related as in the hypothesis of Lemma 2. Clearly T is measurable $\mathscr{F}/\mathscr{F}$.

Although T does not preserve λ, it does preserve Gauss's measure, defined by

$$P(A) = \frac{1}{\log 2} \int_A \frac{dx}{1+x}, \qquad A \in \mathscr{F}. \tag{24.30}$$

In fact, since

$$T^{-1}((0,t) \cap \Omega) = \bigcup_{k=1}^{\infty} \left(\left(\frac{1}{k+t}, \frac{1}{k} \right) \cap \Omega \right),$$

it is enough to verify

$$\int_0^t \frac{dx}{1+x} = \sum_{k=1}^{\infty} \int_{t/(k+1)}^{t/k} \frac{dx}{1+x} = \sum_{k=1}^{\infty} \int_{1/(k+t)}^{1/k} \frac{dx}{1+x}.$$

Gauss's measure is useful because it is preserved by T and has the same sets of measure 0 as Lebesgue measure does.

Proof that T is ergodic. Fix $a_1, \ldots, a_n$, and write ψ_n for $\psi_{a_1 \ldots a_n}$ and Δ_n for $\Delta_{a_1 \ldots a_n}$. Suppose that n is even, so that ψ_n is increasing. If $x \in \Delta_n$, then (since $x = \psi_n(T^n x)$) $s < T^n x < t$ if and only if $\psi_n(s) < x < \psi_n(t)$; and this last condition implies $x \in \Delta_n$. Combined with (24.28) and (24.20), this shows that

$$\lambda(\Delta_n \cap [x: s < T^n x < t]) = \psi_n(t) - \psi_n(s) = \frac{t-s}{(q_n + sq_{n-1})(q_n + tq_{n-1})}.$$

If B is an interval with endpoints s and t, then by (24.29),

$$\lambda(\Delta_n \cap T^{-n}B) = \lambda(\Delta_n)\lambda(B) \frac{q_n(q_n + q_{n-1})}{(q_n + sq_{n-1})(q_n + tq_{n-1})}.$$

A similar argument establishes this for n odd. Since the ratio on the right lies between $\frac{1}{2}$ and 2,

$$\tfrac{1}{2}\lambda(\Delta_n)\lambda(B) \le \lambda(\Delta_n \cap T^{-n}B) \le 2\lambda(\Delta_n)\lambda(B). \tag{24.31}$$

Therefore, (24.10) holds for $\mathscr{P}, \mathscr{F}_0, \mathscr{F}$ as defined above, $A = \Delta_n$, $n_A = n$, $c = \frac{1}{2}$, and λ in the role of P. Thus $T^{-1}C = C$ implies that $\lambda(C)$ is 0 or 1, and since Gauss's measure (24.30) comes from a density, $P(C)$ is 0 or 1 as well. Therefore, *T is an ergodic measure-preserving transformation on* $(\Omega, \mathscr{F}, P)$.

It follows by the ergodic theorem that if f is integrable, then

$$\lim_n \frac{1}{n} \sum_{k=1}^{n} f(T^{k-1}x) = \frac{1}{\log 2} \int_0^1 \frac{f(x)}{1+x}\, dx \tag{24.32}$$

holds almost everywhere. Since the density in (24.30) is bounded away from 0 and ∞, the "integrable" and "almost everywhere" here can refer to P or to λ indifferently.

Taking f to be the indicator of the x-set where $a_1(x) = k$ shows that the asymptotic relative frequency of k among the partial quotients is almost everywhere equal to

$$\frac{1}{\log 2} \int_{1/(k+1)}^{1/k} \frac{dx}{1+x} = \frac{1}{\log 2} \log \frac{(k+1)^2}{k(k+2)}.$$

In particular, the partial quotients are unbounded almost everywhere.

For understanding the accuracy of the continued-fraction algorithm, the magnitude of $a_n(x)$ is less important than that of $q_n(x)$. The key relationship is

$$(24.33)\qquad \frac{1}{q_n(x)(q_n(x)+q_{n+1}(x))} < \left| x - \frac{p_n(x)}{q_n(x)} \right| < \frac{1}{q_n(x)q_{n+1}(x)},$$

which follows from (24.25) and (24.19). Suppose it is shown that

$$(24.34)\qquad \lim_n \frac{1}{n} \log q_n(x) = \frac{\pi^2}{12 \log 2}$$

almost everywhere; (24.33) will then imply that

$$(24.35)\qquad \lim_n \frac{1}{n} \log \left| x - \frac{p_n(x)}{q_n(x)} \right| = -\frac{\pi^2}{6 \log 2}:$$

The discrepancy between x and its nth convergent is almost everywhere of the order $e^{-n\pi^2/(6 \log 2)}$.

To prove (24.34), note first that since $p_{j+1}(x) = q_j(Tx)$ by (24.19), the product $\prod_{k=1}^{n} p_{n-k+1}(T^{k-1}x)/q_{n-k+1}(T^{k-1}x)$ telescopes to $1/q_n(x)$:

$$(24.36)\qquad \log \frac{1}{q_n(x)} = \sum_{k=1}^{n} \log \frac{p_{n-k+1}(T^{k-1}x)}{q_{n-k+1}(T^{k-1}x)}.$$

As observed earlier, $q_n(x) \geq 2^{(n-1)/2}$ for $n \geq 1$. Therefore, by (24.33),

$$\left| \frac{x}{p_n(x)/q_n(x)} - 1 \right| \leq \frac{1}{q_{n+1}(x)} \leq \frac{1}{2^{n/2}}, \quad n \geq 1.$$

Since $|\log(1+s)| \leq 4|s|$ if $|s| \leq 1/\sqrt{2}$,

$$\left| \log x - \log \frac{p_n(x)}{q_n(x)} \right| \leq \frac{4}{2^{n/2}}.$$

Therefore, by (24.36),

$$\left| \sum_{k=1}^{n} \log T^{k-1}x - \log \frac{1}{q_n(x)} \right| \leq \sum_{k=1}^{n} \left| \log T^{k-1}x - \log \frac{p_{n-k+1}(T^{k-1}x)}{q_{n-k+1}(T^{k-1}x)} \right|$$
$$\leq \sum_{i=1}^{\infty} \frac{4}{2^{i/2}} < \infty.$$

By the ergodic theorem, then,†

$$\lim_n \frac{1}{n} \log \frac{1}{q_n(x)} = \frac{1}{\log 2} \int_0^1 \frac{\log x}{1+x}\, dx = \frac{-1}{\log 2} \int_0^1 \log(1+x)\, \frac{dx}{x}$$

$$= \frac{-1}{\log 2} \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)^2} = \frac{-\pi^2}{12 \log 2}.$$

Hence (24.34).

Diophantine Approximation

The fundamental theorem of the measure theory of Diophantine approximation, due to Khinchine, is Theorem 1.5 together with Theorem 1.6. As in Section 1, let $\varphi(q)$ be a positive function of integers and let A_φ be the set of x in $(0, 1)$ such that

$$\left| x - \frac{p}{q} \right| < \frac{1}{q^2 \varphi(q)} \tag{24.37}$$

has infinitely many irreducible solutions p/q. If $\Sigma 1/q\varphi(q)$ converges, then A_φ has Lebesgue measure 0, as was proved in Section 1 (Theorem 1.6). It remains to prove that if φ is nondecreasing and $\Sigma 1/q\varphi(q)$ diverges, then A_φ has Lebesgue measure 1 (Theorem 1.5). It is enough to consider irrational x.

Lemma 3. *For positive α_n, the probability (P or λ) of $[x: a_n(x) > \alpha_n \text{ i.o.}]$ is 0 or 1 as $\Sigma 1/\alpha_n$ converges or diverges.*

PROOF. Let $E_n = [x: a_n(x) > \alpha_n]$. Since $P(E_n) = P[x: a_1(x) > \alpha_n]$ is of the order $1/\alpha_n$, the first Borel–Cantelli lemma settles the convergent case (not needed in the proof of Theorem 1.5).

By (24.31),

$$\lambda(\Delta_n \cap E_{n+1}) \geq \tfrac{1}{2}\lambda(\Delta_n)\lambda[x: a_1(x) > \alpha_{n+1}] \geq \tfrac{1}{2}\lambda(\Delta_n) \frac{1}{\alpha_{n+1}+1}.$$

Taking a union over certain of the Δ_n shows that for $m < n$,

$$\lambda(E_m^c \cap \cdots \cap E_n^c \cap E_{n+1}) \geq \lambda(E_m^c \cap \cdots \cap E_n^c) \frac{1}{2(\alpha_{n+1}+1)}.$$

By induction on n,

$$\lambda(E_m^c \cap \cdots \cap E_n^c) \leq \prod_{k=m}^{n} \left(1 - \frac{1}{2(\alpha_{k+1}+1)}\right)$$

$$\leq \exp\left[- \sum_{k=m}^{n} \frac{1}{2(\alpha_{k+1}+1)} \right],$$

as in the proof of the second Borel–Cantelli lemma. ■

†Integrate by parts over $(\alpha, 1)$ and then let $\alpha \downarrow 0$. For the series, see Problem 26.28. The specific value of the limit in (24.34) is not needed for the application that follows.

PROOF OF THEOREM 1.5. Fix an integer N such that $\log N$ exceeds the limit in (24.34). Then, except on a set of measure 0,

$$q_n(x) < N^n \tag{24.38}$$

holds for all but finitely many n. Since φ is nondecreasing,

$$\sum_{N^n \le q < N^{n+1}} \frac{1}{q\varphi(q)} \le \frac{N}{\varphi(N^n)},$$

and $\sum 1/\varphi(N^n)$ diverges if $\sum 1/q\varphi(q)$ does. By the lemma, outside a set of measure 0, $a_{n+1}(x) \ge \varphi(N^n)$ holds for infinitely many n. If this inequality and (24.38) both hold, then by (24.33) and the assumption that φ is nondecreasing,

$$\left| x - \frac{p_n(x)}{q_n(x)} \right| < \frac{1}{q_n(x)q_{n+1}(x)} \le \frac{1}{a_{n+1}(x)q_n^2(x)}$$

$$\le \frac{1}{\varphi(N^n)q_n^2(x)} \le \frac{1}{\varphi(q_n(x))q_n^2(x)}.$$

But $p_n(x)/q_n(x)$ is irreducible by (24.20). ■

PROBLEMS

24.1. Fix $(\Omega, \mathscr{F})$ and a T measurable $\mathscr{F}/\mathscr{F}$. The probability measures on $(\Omega, \mathscr{F})$ preserved by T form a convex set C. Show that T is ergodic under P if and only if P is an extreme point of C—cannot be represented as a proper convex combination of distinct elements of C.

24.2. Show that T is ergodic if and only if $n^{-1}\sum_{k=1}^{n-1} P(A \cap T^{-k}B) \to P(A)P(B)$ for all A and B (or all A and B in a π-system generating $\mathscr{F}$).

24.3. ↑ The transformation T is *mixing* if

$$P(A \cap T^{-n}B) \to P(A)P(B) \tag{24.39}$$

for all A and B.

(a) Show that mixing implies ergodicity.

(b) Show that T is mixing if (24.39) holds for all A and B in a π-system generating $\mathscr{F}$.

(c) Show that the Bernoulli shift is mixing.

(d) Show that a cyclic permutation is ergodic but not mixing.

(e) Show that if c is not a root of unity, then the rotation (Example 24.4) is ergodic but not mixing.

24.4. ↑ Write $T^{-n}\mathscr{F}=[T^{-n}A\colon A\in\mathscr{F}]$, and call the σ-field $\mathscr{F}_\infty=\bigcap_{n=1}^{\infty}T^{-n}\mathscr{F}$ *trivial* if every set in it has probability either 0 or 1. (If T is invertible, $\mathscr{F}_\infty$ is $\mathscr{F}$ and hence is trivial only in uninteresting cases.)

(a) Show that if $\mathscr{F}_\infty$ is trivial, then T is ergodic. (A cyclic permutation is ergodic even though $\mathscr{F}_\infty$ is not trivial.)

(b) Show that if the hypotheses of Lemma 2 are satisfied, then $\mathscr{F}_\infty$ is trivial.

(c) It can be shown by martingale theory that if $\mathscr{F}_\infty$ is trivial, then T is mixing; see Problem 35.20. Reconsider Problem 24.3(c).

24.5. 8.35 24.4 ↑ **(a)** Show that the shift corresponding to an irreducible, aperiodic Markov chain is mixing. Do this first by Problem 8.35, then by Problem 24.4(b), (c).

(b) Show that if the chain is irreducible but has period greater than 1, then the shift is ergodic but not mixing.

(c) Suppose the state space splits into two closed, disjoint, nonempty subsets, and that the initial distribution (stationary) gives positive weight to each. Show that the corresponding shift is not ergodic.

24.6. Show that if T is ergodic and if f is nonnegative and $E[f]=\infty$, then $n^{-1}\sum_{k=1}^{n}f(T^{k-1}\omega)\to\infty$ with probability 1.

24.7. 24.3 ↑ Suppose that $P_0(A)=\int_A\delta\,dP$ for all A ($\delta\geq 0$) and that T is mixing with respect to P (T need not preserve P_0). Use (21.9) to prove

$$P_0(T^{-n}A)=\int_{T^{-n}A}\delta\,dP\to P(A).$$

24.8. 24.6 ↑ **(a)** Show that

$$\frac{1}{n}\sum_{k=1}^{n}a_k(x)\to\infty$$

and

$$\sqrt[n]{a_1(x)\cdots a_n(x)}\to\prod_{k=1}^{\infty}\left(1+\frac{1}{k^2+2k}\right)^{(\log k)/(\log 2)}$$

almost everywhere.

(b) Show that

$$\frac{1}{n}\log\left(q_n(x)\left|x-\frac{p_n(x)}{q_n(x)}\right|\right)\to-\frac{\pi^2}{12\log 2}.$$

24.9. 24.4 24.7 ↑ **(a)** Show that the continued-fraction transformation is mixing.

(b) Show that

$$\lambda[x\colon T^n x\leq t]\to\frac{\log(1+t)}{\log 2},\qquad 0\leq t\leq 1.$$

CHAPTER 5

Convergence of Distributions

SECTION 25. WEAK CONVERGENCE

Many of the best-known theorems in probability have to do with the asymptotic behavior of distributions. This chapter covers both general methods for deriving such theorems and specific applications. The present section concerns the general limit theory for distributions on the real line, and the methods of proof use in an essential way the order structure of the line. For the theory in R^k, see Section 29.

Definitions

Distribution functions F_n were defined in Section 14 to *converge weakly* to the distribution function F if

$$\lim_n F_n(x) = F(x) \tag{25.1}$$

for every continuity point x of F; this is expressed by writing $F_n \Rightarrow F$. Examples 14.1, 14.2, and 14.3 illustrate this concept in connection with the asymptotic distribution of maxima. Example 14.4 shows the point of allowing (25.1) to fail if F is discontinuous at x; see also Example 25.4. Theorem 25.8 and Example 25.9 show why this exemption is essential to a useful theory.

If μ_n and μ are the probability measures on $(R^1, \mathscr{R}^1)$ corresponding to F_n and F, then $F_n \Rightarrow F$ if and only if

$$\lim_n \mu_n(A) = \mu(A) \tag{25.2}$$

for every A of the form $A = (-\infty, x]$ for which $\mu\{x\} = 0$—see (20.5). In this case the distributions themselves are said to converge weakly, which is expressed by writing $\mu_n \Rightarrow \mu$. Thus $F_n \Rightarrow F$ and $\mu_n \Rightarrow \mu$ are only different expressions of the same fact. From weak convergence it follows that (25.2) holds for many sets A besides half-infinite intervals; see Theorem 25.8.

Example 25.1. Let F_n be the distribution function corresponding to a unit mass at n: $F_n = I_{[n,\infty)}$. Then $\lim_n F_n(x) = 0$ for every x, so that (25.1) is satisfied if $F(x) \equiv 0$. But $F_n \Rightarrow F$ does not hold, because F is not a distribution function. Weak convergence is defined in this section only for functions F_n and F that rise from 0 at $-\infty$ to 1 at $+\infty$—that is, it is defined only for *probability* measures μ_n and μ.[†] ■

Example 25.2. *Poisson approximation to the binomial.* Let μ_n be the binomial distribution (20.6) for $p = \lambda/n$ and let μ be the Poisson distribution (20.7). For nonnegative integers k,

$$\mu_n\{k\} = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$
$$= \frac{\lambda^k(1-\lambda/n)^n}{k!} \times \frac{1}{(1-\lambda/n)^k}\prod_{i=0}^{k-1}\left(1-\frac{i}{n}\right)$$

if $n \geq k$. As $n \to \infty$ the second factor on the right goes to 1 for fixed k, and so $\mu_n\{k\} \to \mu\{k\}$; this is a special case of Theorem 23.2. By the series form of Scheffé's theorem (the corollary to Theorem 16.12), (25.2) holds for every set A of nonnegative integers. Since the nonnegative integers support μ and the μ_n, (25.2) even holds for every linear Borel set A. Certainly μ_n converges weakly to μ in this case. ■

Example 25.3. Let μ_n correspond to a mass of n^{-1} at each point k/n, $k = 0, 1, \ldots, n-1$; let μ be Lebesgue measure confined to the unit interval. The corresponding distribution functions satisfy $F_n(x) = (\lfloor nx \rfloor + 1)/n \to F(x)$ for $0 \leq x < 1$, and so $F_n \Rightarrow F$. In this case (25.1) holds for every x, but (25.2) does not, as in the preceding example, hold for every Borel set A: if A is the set of rationals, then $\mu_n(A) = 1$ does not converge to $\mu(A) = 0$. Despite this, μ_n does converge weakly to μ. ■

Example 25.4. If μ_n is a unit mass at x_n and μ is a unit mass at x, then $\mu_n \Rightarrow \mu$ if and only if $x_n \to x$. If $x_n > x$ for infinitely many n, then (25.1) fails at the discontinuity point of F. ■

Uniform Distribution Modulo 1*

For a sequence $x_1, x_2, \ldots$ of real numbers, consider the corresponding sequence of their fractional parts $\{x_n\} = x_n - \lfloor x_n \rfloor$. For each n, define a probability measure μ_n by

$$\mu_n(A) = \frac{1}{n}\#[k: 1 \leq k \leq n, \{x_k\} \in A]; \tag{25.3}$$

[†]There is (see Section 28) a related notion of *vague* convergence in which μ may be *defective* in the sense that $\mu(R^1) < 1$. Weak convergence is in this context sometimes called *complete convergence*.

*This topic, which requires ergodic theory, may be omitted.

μ_n has mass n^{-1} at the points $\{x_1\},\dots,\{x_n\}$, and if several of these points coincide, the masses add. The problem is to find the weak limit of $\{\mu_n\}$ in number-theoretically interesting cases.

If the μ_n defined by (25.3) converge weakly to Lebesgue measure restricted to the unit interval, the sequence $x_1, x_2, \dots$ is said to be *uniformly distributed modulo* 1. In this case every subinterval has asymptotically its proportional share of the points $\{x_n\}$; by Theorem 25.8 below, the same is then true of every subset whose boundary has Lebesgue measure 0.

Theorem 25.1. *For θ irrational, $\theta, 2\theta, 3\theta, \dots$ is uniformly distributed modulo* 1.

PROOF. Since $\{n\theta\} = \{n\{\theta\}\}$, θ can be assumed to lie in [0,1]. As in Example 24.4, map [0, 1) to the unit circle in the complex plane by $\phi(x) = e^{2\pi i x}$. If θ is irrational, then $c = \phi(\theta)$ is not a root of unity, and so (p. 000) $T\omega = c\omega$ defines an ergodic transformation with respect to circular Lebesgue measure P. Let $\mathscr{I}$ be the class of open arcs with endpoints in some fixed countable, dense set. By the ergodic theorem, the orbit $\{T^n\omega\}$ of almost every ω enters every I in $\mathscr{I}$ with asymptotic relative frequency $P(I)$. Fix such an ω. If $I_1 \subset J \subset I_2$, where J is a closed arc and I_1, I_2 are in $\mathscr{I}$, then the upper and lower limits of $n^{-1}\sum_{k=1}^{n} I_J(T^{k-1}\omega)$ are between $P(I_1)$ and $P(I_2)$, and therefore the limit exists and equals $P(J)$. Since the orbits and the arcs are rotations of one another, *every* orbit enters *every* closed arc J with frequency $P(J)$. This is true in particular of the orbit $\{c^n\}$ of 1.

Now carry all this back to [0, 1) by ϕ^{-1}: For every x in [0, 1), $\{n\theta\} = \phi^{-1}(c^n)$ lies in $[0, x]$ with asymptotic relative frequency x. ■

For a simple proof by Fourier series, see Example 26.3.

Convergence in Distribution

Let X_n and X be random variables with respective distribution functions F_n and F. If $F_n \Rightarrow F$, then X_n is said to *converge in distribution* or *in law* to X, written $X_n \Rightarrow X$. This dual use of the double arrow will cause no confusion. Because of the defining conditions (25.1) and (25.2), $X_n \Rightarrow X$ if and only if

$$\lim_n P[X_n \le x] = P[X \le x] \tag{25.4}$$

for every x such that $P[X = x] = 0$.

Example 25.5. Let $X_1, X_2, \dots$ be independent random variables, each with the exponential distribution: $P[X_n \ge x] = e^{-\alpha x}$, $x \ge 0$. Put $M_n = \max\{X_1, \dots, X_n\}$ and $b_n = \alpha^{-1} \log n$. The relation (14.9), established in Example 14.1, can be restated as $P[M_n - b_n \le x] \to e^{-e^{-\alpha x}}$. If X is any random variable with distribution function $e^{-e^{-\alpha x}}$, this can be written $M_n - b_n \Rightarrow X$. ■

One is usually interested in proving weak convergence of the distributions of some given sequence of random variables, such as the $M_n - b_n$ in this example, and the result is often most clearly expressed in terms of the random variables themselves rather than in terms of their distributions or

distribution functions. Although the $M_n - b_n$ here arise naturally from the problem at hand, the random variable X is simply constructed to make it possible to express the asymptotic relation compactly by $M_n - b_n \Rightarrow X$. Recall that by Theorem 14.1 there does exist a random variable for any prescribed distribution.

Example 25.6. For each n, let Ω_n be the space of n-tuples of 0's and 1's, let $\mathscr{F}_n$ consist of all subsets of Ω_n, and let P_n assign probability $(\lambda/n)^k(1-\lambda/n)^{n-k}$ to each ω consisting of k 1's and $n-k$ 0's. Let $X_n(\omega)$ be the number of 1's in ω; then X_n, a random variable on $(\Omega_n, \mathscr{F}_n, P_n)$, represents the number of successes in n Bernoulli trials having probability λ/n of success at each.

Let X be a random variable, on some $(\Omega, \mathscr{F}, P)$, having the Poisson distribution with parameter λ. According to Example 25.2, $X_n \Rightarrow X$. ■

As this example shows, the random variables X_n may be defined on entirely different probability spaces. To allow for this possibility, the P on the left in (25.4) really should be written P_n. Suppressing the n causes no confusion if it is understood that P refers to whatever probability space it is that X_n is defined on; the underlying probability space enters into the definition only via the distribution μ_n it induces on the line. Any instance of $F_n \Rightarrow F$ or of $\mu_n \Rightarrow \mu$ can be rewritten in terms of convergence in distribution: There exist random variables X_n and X (on some probability spaces) with distribution functions F_n and F, and $F_n \Rightarrow F$ and $X_n \Rightarrow X$ express the same fact.

Convergence in Probability

Suppose that $X, X_1, X_2, \ldots$ are random variables all defined on the same probability space $(\Omega, \mathscr{F}, P)$. If $X_n \to X$ with probability 1, then $P[|X_n - X| \geq \epsilon \text{ i.o.}] = 0$ for $\epsilon > 0$, and hence

$$\lim_n P\left[|X_n - X| > \epsilon\right] = 0 \tag{25.5}$$

by Theorem 4.1. Thus there is *convergence in probability* $X_n \to_P X$; see Theorems 5.2 and 20.5.

Suppose that (25.5) holds for each positive ϵ. Now $P[X \leq x - \epsilon] - P[|X_n - X| \geq \epsilon] \leq P[X_n \leq x] \leq P[X \leq x + \epsilon] + P[|X_n - x| \geq \epsilon]$; letting n tend to ∞ and then letting ϵ tend to 0 shows that $P[X < x] \leq \liminf_n P[X_n \leq x] \leq \limsup_n P[X_n \leq x] \leq P[X \leq x]$. Thus $P[X_n \leq x] \to P[X \leq x]$ if $P[X = x] = 0$, and so $X_n \Rightarrow X$:

Theorem 25.2. *Suppose that X_n and X are random variables on the same probability space. If $X_n \to X$ with probability 1, then $X_n \to_P X$. If $X_n \to_P X$, then $X_n \Rightarrow X$.*

Of the two implications in this theorem, neither converse holds. Because of Example 5.4, convergence in probability does not imply convergence with probability 1. Neither does convergence in distribution imply convergence in probability: if X and Y are independent and assume the values 0 and 1 with probability $\frac{1}{2}$ each, and if $X_n = Y$, then $X_n \Rightarrow X$, but $X_n \to_P X$ cannot hold because $P[|X - Y|] = 1 = \frac{1}{2}$. What is more, (25.5) is impossible if X and the X_n are defined on different probability spaces, as may happen in the case of convergence in distribution.

Although (25.5) in general makes no sense unless X and the X_n are defined on the same probability space, suppose that X is replaced by a constant real number a—that is, suppose that $X(\omega) \equiv a$. Then (25.5) becomes

$$\lim_n P[|X_n - a| \geq \epsilon] = 0, \tag{25.6}$$

and this condition makes sense even if the space of X_n does vary with n. Now a can be regarded as a random variable (on any probability space at all), and it is easy to show that (25.6) implies that $X_n \Rightarrow a$: Put $\epsilon = |x - a|$; if $x > a$, then $P[X_n \leq x] \geq P[|X_n - a| < \epsilon] \to 1$, and if $x < a$, then $P[X_n \leq x] \leq P[|X_n - a| \geq \epsilon] \to 0$. If a is regarded as a random variable, its distribution function is 0 for $x < a$ and 1 for $x \geq a$. Thus (25.6) implies that the distribution function of X_n converges weakly to that of a.

Suppose, on the other hand, that $X_n \Rightarrow a$. Then $P[|X_n - a| > \epsilon] \leq P[X_n \leq a - \epsilon] + 1 - P[X_n \leq a + \epsilon] \to 0$, so that (25.6) holds:

Theorem 25.3. *The condition* (25.6) *holds for all positive* ϵ *if and only if* $X_n \Rightarrow a$, *that is, if and only if*

$$\lim_n P[X_n \leq x] = \begin{cases} 0 & \text{if } x < a, \\ 1 & \text{if } x > a. \end{cases}$$

If (25.6) holds for all positive ϵ, X_n may be said to *converge to a in probability*. As this does not require that the X_n be defined on the same space, it is not really a special case of convergence in probability as defined by (25.5). Convergence in probability in this new sense will be denoted $X_n \Rightarrow a$, in accordance with the theorem just proved.

Example 14.4 restates the weak law of large numbers in terms of this concept. Indeed, if $X_1, X_2, \ldots$ are independent, identically distributed random variables with finite mean m, and if $S_n = X_1 + \cdots + X_n$, the weak law of large numbers is the assertion $n^{-1}S_n \Rightarrow m$. Example 6.3 provides another illustration: If S_n is the number of cycles in a random permutation on n letters, then $S_n/\log n \Rightarrow 1$.

Example 25.7. Suppose that $X_n \Rightarrow X$ and $\delta_n \to 0$. Given ϵ and η, choose x so that $P[|X| \geq x] < \eta$ and $P[X = \pm x] = 0$, and then choose n_0 so that $n \geq n_0$ implies that $|\delta_n| < \epsilon/x$ and $|P[X_n \leq y] - P[X \leq y]| < \eta$ for $y = \pm x$. Then $P[|\delta_n X_n| \geq \epsilon] < 3\eta$ for $n \geq n_0$. Thus $X_n \Rightarrow X$ *and* $\delta_n \to 0$ *imply that* $\delta_n X_n \Rightarrow 0$, a restatement of Lemma 2 of Section 14 (p. 193). ■

The asymptotic properties of a random variable should remain unaffected if it is altered by the addition of a random variable that goes to 0 in probability. Let (X_n, Y_n) be a two-dimensional random vector.

Theorem 25.4. *If* $X_n \Rightarrow X$ *and* $X_n - Y_n \Rightarrow 0$, *then* $Y_n \Rightarrow X$.

PROOF. Suppose that $y' < x < y''$ and $P[X = y'] = P[X = y''] = 0$. If $y' < x - \epsilon < x < x + \epsilon < y''$, then

$$P[X_n \leq y'] - P[|X_n - Y_n| \geq \epsilon] \leq P[Y_n \leq x] \leq P[X_n \leq y''] + P[|X_n - Y_n| \geq \epsilon]. \tag{25.7}$$

Since $X_n \Rightarrow X$, letting $n \to \infty$ gives

$$P[X \leq y'] \leq \liminf_n P[Y_n \leq x] \leq \limsup_n P[Y_n \leq x] \leq P[X \leq y'']. \tag{25.8}$$

Since $P[X = y] = 0$ for all but countably many y, if $P[X = x] = 0$, then y' and y'' can further be chosen so that $P[X \leq y']$ and $P[X \leq y'']$ are arbitrarily near $P[X \leq x]$; hence $P[Y_n \leq x] \to P[X \leq x]$. ■

Theorem 25.4 has a useful extension. Suppose that $(X_n^{(u)}, Y_n)$ is a two-dimensional random vector.

Theorem 25.5. *If, for each* u, $X_n^{(u)} \Rightarrow X^{(u)}$ *as* $n \to \infty$, *if* $X^{(u)} \Rightarrow X$ *as* $u \to \infty$, *and if*

$$\lim_u \limsup_n P\left[|X_n^{(u)} - Y_n| \geq \epsilon\right] = 0 \tag{25.9}$$

for positive ϵ, *then* $Y_n \Rightarrow X$.

PROOF. Replace X_n by $X_n^{(u)}$ in (25.7). If $P[X = y'] = 0 \equiv P[X^{(u)} = y']$ and $P[X = y''] = 0 \equiv P[X^{(u)} = y'']$, letting $n \to \infty$ and then $u \to \infty$ gives (25.8) once again. Since $P[X = y] = 0 \equiv P[X^{(u)} = y]$ for all but countably many y, the proof can be completed as before. ■

Fundamental Theorems

Some of the fundamental properties of weak convergence were established in Section 14. It was shown there that a sequence cannot have two distinct weak limits: *If* $F_n \Rightarrow F$ *and* $F_n \Rightarrow G$, *then* $F = G$. The proof is simple: The hypothesis implies that F and G agree at their common points of continuity, hence at all but countably many points, and hence by right continuity at all points. Another simple fact is this: *If* $\lim_n F_n(d) = F(d)$ *for* d *in a set* D *dense in* R^1, *then* $F_n \Rightarrow F$. Indeed, if F is continuous at x, there are in D points d' and d'' such that $d' < x < d''$ and $F(d'') - F(d') < \epsilon$, and it follows that the limits superior and inferior of $F_n(x)$ are within ϵ of $F(x)$.

For any probability measure on $(R^1, \mathscr{R}^1)$ there is on some probability space a random variable having that measure as its distribution. Therefore, for probability measures satisfying $\mu_n \Rightarrow \mu$, there exist random variables Y_n and Y having these measures as distributions and satisfying $Y_n \Rightarrow Y$. According to the following theorem, the Y_n and Y can be constructed on the same probability space, and even in such a way that $Y_n(\omega) \to Y(\omega)$ for every ω—a condition much stronger than $Y_n \Rightarrow Y$. This result, *Skorohod's theorem*, makes possible very simple and transparent proofs of many important facts.

Theorem 25.6. *Suppose that* μ_n *and* μ *are probability measures on* $(R^1, \mathscr{R}^1)$ *and* $\mu_n \Rightarrow \mu$. *There exist random variables* Y_n *and* Y *on a common probability space* $(\Omega, \mathscr{F}, P)$ *such that* Y_n *has distribution* μ_n, Y *has distribution* μ, *and* $Y_n(\omega) \to Y(\omega)$ *for each* ω.

PROOF. For the probability space $(\Omega, \mathscr{F}, P)$, take $\Omega = (0, 1)$, let $\mathscr{F}$ consist of the Borel subsets of $(0, 1)$, and for $P(A)$ take the Lebesgue measure of A.

The construction is related to that in the proofs of Theorems 14.1 and 20.4. Consider the distribution functions F_n and F corresponding to μ_n and μ. For $0 < \omega < 1$, put $Y_n(\omega) = \inf[x: \omega \le F_n(x)]$ and $Y(\omega) = \inf[x: \omega \le F(x)]$. Since $\omega \le F_n(x)$ if and only if $Y_n(\omega) \le x$ (see the argument following (14.5)), $P[\omega: Y_n(\omega) \le x] = P[\omega: \omega \le F_n(x)] = F_n(x)$. Thus Y_n has distribution function F_n; similarly, Y has distribution function F.

It remains to show that $Y_n(\omega) \to Y(\omega)$. The idea is that Y_n and Y are essentially inverse functions to F_n and F; if the direct functions converge, so must the inverses.

Suppose that $0 < \omega < 1$. Given ϵ, choose x so that $Y(\omega) - \epsilon < x < Y(\omega)$ and $\mu\{x\} = 0$. Then $F(x) < \omega$; $F_n(x) \to F(x)$ now implies that, for n large enough, $F_n(x) < \omega$ and hence $Y(\omega) - \epsilon < x < Y_n(\omega)$. Thus $\liminf_n Y_n(\omega) \ge Y(\omega)$. If $\omega < \omega'$ and ϵ is positive, choose a y for which $Y(\omega') < y < Y(\omega') + \epsilon$ and $\mu\{y\} = 0$. Now $\omega < \omega' \le F(Y(\omega')) \le F(y)$, and so, for n large enough, $\omega \le F_n(y)$ and hence $Y_n(\omega) \le y < Y(\omega') + \epsilon$. Thus $\limsup_n Y_n(\omega) \le Y(\omega')$ if $\omega < \omega'$. Therefore, $Y_n(\omega) \to Y(\omega)$ if Y is continuous at ω.

Since Y is nondecreasing on $(0,1)$, it has at most countably many discontinuities. At discontinuity points ω of Y, redefine $Y_n(\omega) = Y(\omega) = 0$. With this change, $Y_n(\omega) \to Y(\omega)$ for every ω. Since Y and the Y_n have been altered only on a set of Lebesgue measure 0, their distributions are still μ_n and μ. ■

Note that this proof uses the order structure of the real line in an essential way. The proof of the corresponding result in R^k is more complicated.

The following *mapping theorem* is of very frequent use.

Theorem 25.7. *Suppose that h: $R^1 \to R^1$ is measurable and that the set D_h of its discontinuities is measurable.*[†] *If $\mu_n \Rightarrow \mu$ and $\mu(D_h) = 0$, then $\mu_n h^{-1} \Rightarrow \mu h^{-1}$.*

Recall (see (13.7)) that μh^{-1} has value $\mu(h^{-1}A)$ at A.

PROOF. Consider the random variables Y_n and Y of Theorem 25.6. Since $Y_n(\omega) \to Y(\omega)$, if $Y(\omega) \notin D_h$ then $h(Y_n(\omega)) \to h(Y(\omega))$. Since $P[\omega\colon Y(\omega) \in D_h] = \mu(D_h) = 0$, it follows that $h(Y_n(\omega)) \to h(Y(\omega))$ with probability 1. Hence $h(Y_n) \Rightarrow h(Y)$ by Theorem 25.2. Since $P[h(Y) \in A] = P[Y \in h^{-1}A] = \mu(h^{-1}A)$, $h(Y)$ has distribution μh^{-1}; similarly, $h(Y_n)$ has distribution $\mu_n h^{-1}$. Thus $h(Y_n) \Rightarrow h(Y)$ is the same thing as $\mu_n h^{-1} \Rightarrow \mu h^{-1}$. ■

Because of the definition of convergence in distribution, this result has an equivalent statement in terms of random variables:

Corollary 1. *If $X_n \Rightarrow X$ and $P[X \in D_h] = 0$, then $h(X_n) \Rightarrow h(X)$.*

Take $X \equiv a$:

Corollary 2. *If $X_n \Rightarrow a$ and h is continuous at a, then $h(X_n) \Rightarrow h(a)$.*

Example 25.8. From $X_n \Rightarrow X$ it follows directly by the theorem that $aX_n + b \Rightarrow aX + b$. Suppose also that $a_n \to a$ and $b_n \to b$. Then $(a_n - a)X_n \Rightarrow 0$ by Example 25.7, and so $(a_n X_n + b_n) - (aX_n + b) \Rightarrow 0$. And now $a_n X_n + b_n \Rightarrow aX + b$ follows by Theorem 25.4: *If $X_n \Rightarrow X$, $a_n \to a$, and $b_n \to b$, then $a_n X_n + b_n \Rightarrow aX + b$.* This fact was stated and proved differently in Section 14 —see Lemma 1 on p. 193. ■

By definition, $\mu_n \Rightarrow \mu$ means that the corresponding distribution functions converge weakly. The following theorem characterizes weak convergence

[†]That D_h lies in $\mathscr{R}^1$ is generally obvious in applications. In point of fact, it always holds (even if h is not measurable): Let $A(\epsilon, \delta)$ be the set of x for which there exist y and z such that $|x - y| < \delta$, $|x - z| < \delta$, and $|h(y) - h(z)| \geq \epsilon$. Then $A(\epsilon, \delta)$ is open and $D_h = \bigcup_\epsilon \bigcap_\delta A(\epsilon, \delta)$, where ϵ and δ range over the positive rationals.

without reference to distribution functions. The boundary ∂A of A consists of the points that are limits of sequences in A and are also limits of sequences in A^c; alternatively, ∂A is the closure of A minus its interior. A set A is a *μ-continuity set* if it is a Borel set and $\mu(\partial A) = 0$.

Theorem 25.8. *The following three conditions are equivalent.*

(i) $\mu_n \Rightarrow \mu$;

(ii) $\int f d\mu_n \to \int f d\mu$ *for every bounded, continuous real function* f;

(iii) $\mu_n(A) \to \mu(A)$ *for every* μ*-continuity set* A.

PROOF. Suppose that $\mu_n \Rightarrow \mu$, and consider the random variables Y_n and Y of Theorem 25.6. Suppose that f is a bounded function such that $\mu(D_f) = 0$, where D_f is the set of points of discontinuity of F. From $P[Y \in D_f] = \mu(D_f) = 0$ it follows that $f(Y_n) \to f(Y)$ with probability 1, and so by change of variable (see (21.1)) and the bounded convergence theorem, $\int f d\mu_n = E[f(Y_n)] \to E[f(Y)] = \int f d\mu$. Thus $\mu_n \Rightarrow \mu$ and $\mu(D_f) = 0$ together imply that $\int f d\mu_n \to \int f d\mu$ if f is bounded. In particular, (i) implies (ii). Further, if $f = I_A$, then $D_f = \partial A$, and from $\mu(\partial A) = 0$ and $\mu_n \Rightarrow \mu$ follows $\mu_n(A) = \int f d\mu_n \to \int f d\mu = \mu(A)$. Thus (i) also implies (iii).

Since $\partial(-\infty, x] = \{x\}$, obviously (iii) implies (i). It therefore remains only to deduce $\mu_n \Rightarrow \mu$ from (ii). Consider the corresponding distribution functions. Suppose that $x < y$, and let $f(t)$ be 1 for $t \le x$, 0 for $t \ge y$, and interpolate linearly on $[x, y]$: $f(t) = (y - t)/(y - x)$ for $x \le t \le y$. Since $F_n(x) \le \int f d\mu_n$ and $\int f d\mu \le F(y)$, it follows from (ii) that $\limsup_n F_n(x) \le F(y)$; letting $y \downarrow x$ shows that $\limsup_n F_n(x) \le F(x)$. Similarly, $F(u) \le \liminf_n F_n(x)$ for $u < x$ and hence $F(x-) \le \liminf_n F_n(x)$. This implies convergence at continuity points. ■

The function f in this last part of the proof is uniformly continuous. Hence $\mu_n \Rightarrow \mu$ follows if $\int f d\mu_n \to \int f d\mu$ for every bounded and *uniformly* continuous f.

Example 25.9. The distributions in Example 25.3 satisfy $\mu_n \Rightarrow \mu$, but $\mu_n(A)$ does not converge to $\mu(A)$ if A is the set of rationals. Hence this A cannot be a μ-continuity set; in fact, of course, $\partial A = R^1$. ■

The concept of weak convergence would be nearly useless if (25.2) were not allowed to fail when $\mu(\partial A) > 0$. Since $F(x) - F(x-) = \mu\{x\} = \mu(\partial(-\infty, x])$, it is therefore natural in the original definition to allow (25.1) to fail when x is not a continuity point of F.

Helly's Theorem

One of the most frequently used results in analysis is the *Helly selection theorem*:

Theorem 25.9. *For every sequence $\{F_n\}$ of distribution functions there exists a subsequence $\{F_{n_k}\}$ and a nondecreasing, right-continuous function F such that $\lim_k F_{n_k}(x) = F(x)$ at continuity points x of F.*

PROOF. An application of the diagonal method [A14] gives a sequence $\{n_k\}$ of integers along which the limit $G(r) = \lim_k F_{n_k}(r)$ exists for every rational r. Define $F(x) = \inf[G(r): x < r]$. Clearly F is nondecreasing.

To each x and ϵ there is an r for which $x < r$ and $G(r) < F(x) + \epsilon$. If $x \le y < r$, then $F(y) \le G(r) < F(x) + \epsilon$. Hence F is continuous from the right.

If F is continuous at x, choose $y < x$ so that $F(x) - \epsilon < F(y)$; now choose rational r and s so that $y < r < x < s$ and $G(s) < F(x) + \epsilon$. From $F(x) - \epsilon < G(r) \le G(s) < F(x) + \epsilon$ and $F_n(r) \le F_n(x) \le F_n(s)$ it follows that as k goes to infinity $F_{n_k}(x)$ has limits superior and inferior within ϵ of $F(x)$. ■

The F in this theorem necessarily satisfies $0 \le F(x) \le 1$. But F need not be a distribution function: if F_n has a unit jump at n, for example, $F(x) \equiv 0$ is the only possibility. It is important to have a condition which ensures that for some subsequence the limit F is a distribution function.

A sequence of probability measures μ_n on $(R^1, \mathscr{R}^1)$ is said to be *tight* if for each ϵ there exists a finite interval $(a, b]$ such that $\mu_n(a, b] > 1 - \epsilon$ for all n. In terms of the corresponding distribution functions F_n, the condition is that for each ϵ there exist x and y such that $F_n(x) < \epsilon$ and $F_n(y) > 1 - \epsilon$ for all n. If μ_n is a unit mass at n, $\{\mu_n\}$ is not tight in this sense—the mass of μ_n "escapes to infinity." Tightness is a condition preventing this escape of mass.

Theorem 25.10. *Tightness is a necessary and sufficient condition that for every subsequence $\{\mu_{n_k}\}$ there exist a further subsequence $\{\mu_{n_{k(j)}}\}$ and a probability measure μ such that $\mu_{n_{k(j)}} \Rightarrow \mu$ as $j \to \infty$.*

Only the sufficiency of the condition in this theorem is used in what follows.

PROOF. *Sufficiency.* Apply Helly's theorem to the subsequence $\{F_{n_k}\}$ of corresponding distribution functions. There exists a further subsequence $\{F_{n_{k(j)}}\}$ such that $\lim_j F_{n_{k(j)}}(x) = F(x)$ at continuity points of F, where F is nondecreasing and right-continuous. There exists by Theorem 12.4 a measure μ on $(R^1, \mathscr{R}^1)$ such that $\mu(a, b] = F(b) - F(a)$. Given ϵ, choose a and b so that $\mu_n(a, b] > 1 - \epsilon$ for all n, which is possible by tightness. By decreasing a

and increasing b, one can ensure that they are continuity points of F. But then $\mu(a,b] \geq 1-\epsilon$. Therefore, μ is a *probability* measure, and of course $\mu_{n_{k(j)}} \Rightarrow \mu$.

Necessity. If $\{\mu_n\}$ is not tight, there exists a positive ϵ such that for each finite interval $(a,b]$, $\mu_n(a,b] \leq 1-\epsilon$ for some n. Choose n_k so that $\mu_{n_k}(-k,k] \leq 1-\epsilon$. Suppose that some subsequence $\{\mu_{n_{k(j)}}\}$ of $\{\mu_{n_k}\}$ were to converge weakly to some probability measure μ. Choose $(a,b]$ so that $\mu\{a\} = \mu\{b\} = 0$ and $\mu(a,b] > 1-\epsilon$. For large enough j, $(a,b] \subset (-k(j),k(j)]$, and so $1-\epsilon \geq \mu_{n_{k(j)}}(-k(j),k(j)] \geq \mu_{n_{k(j)}}(a,b] \to \mu(a,b]$. Thus $\mu(a,b] \leq 1-\epsilon$, a contradiction. ■

Corollary. *If $\{\mu_n\}$ is a tight sequence of probability measures, and if each subsequence that converges weakly at all converges weakly to the probability measure μ, then $\mu_n \Rightarrow \mu$.*

PROOF. By the theorem, each subsequence $\{\mu_{n_k}\}$ contains a further subsequence $\{\mu_{n_{k(j)}}\}$ converging weakly ($j \to \infty$) to some limit, and that limit must by hypothesis be μ. Thus every subsequence $\{\mu_{n_k}\}$ contains a further subsequence $\{\mu_{n_{k(j)}}\}$ converging weakly to μ.

Suppose that $\mu_n \Rightarrow \mu$ is false. Then there exists some x such that $\mu\{x\} = 0$ but $\mu_n(-\infty,x]$ does not converge to $\mu(-\infty,x]$. But then there exists a positive ϵ such that $|\mu_{n_k}(-\infty,x] - \mu(-\infty,x]| \geq \epsilon$ for an infinite sequence $\{n_k\}$ of integers, and no subsequence of $\{\mu_{n_k}\}$ can converge weakly to μ. This contradiction shows that $\mu_n \Rightarrow \mu$. ■

If μ_n is a unit mass at x_n, then $\{\mu_n\}$ is tight if and only if $\{x_n\}$ is bounded. The theorem above and its corollary reduce in this case to standard facts about real line; see Example 25.4 and A10: tightness of sequences of probability measures is analogous to boundedness of sequences of real numbers.

Example 25.10. Let μ_n be the normal distribution with mean m_n and variance σ_n^2. If m_n and σ_n^2 are bounded, then the second moment of μ_n is bounded, and it follows by Markov's inequality (21.12) that $\{\mu_n\}$ is tight. The conclusion of Theorem 25.10 can also be checked directly: If $\{n_{k(j)}\}$ is chosen so that $\lim_j m_{n_{k(j)}} = m$ and $\lim_j \sigma^2_{n_{k(j)}} = \sigma^2$, then $\mu_{n_{k(j)}} \Rightarrow \mu$, where μ is normal with mean m and variance σ^2 (a unit mass at m if $\sigma^2 = 0$).

If $m_n > b$, then $\mu_n(b,\infty) \geq \frac{1}{2}$; if $m_n < a$, then $\mu_n(-\infty,a] \geq \frac{1}{2}$. Hence $\{\mu_n\}$ cannot be tight if m_n is unbounded. If m_n is bounded, say by K, then $\mu_n(-\infty,a] \geq \nu(-\infty,(a-K)\sigma_n^{-1}]$, where ν is the standard normal distribution. If σ_n is unbounded, then $\nu(-\infty,(a-K)\sigma_n^{-1}] \to \frac{1}{2}$ along some subsequence, and $\{\mu_n\}$ cannot be tight. Thus a sequence of normal distributions is tight if and only if the means and variances are bounded. ■

Integration to the Limit

Theorem 25.11. *If $X_n \Rightarrow X$, then $E[|X|] \le \liminf_n E[|X_n|]$.*

PROOF. Apply Skorohod's Theorem 25.6 to the distributions of X_n and X: There exist on a common probability space random variables Y_n and Y such that $Y = \lim_n Y_n$ with probability 1, Y_n has the distribution of X_n, and Y has the distribution of X. By Fatou's lemma, $E[|Y|] \le \liminf_n E[|Y_n|]$. Since $|X|$ and $|Y|$ have the same distribution, they have the same expected value (see (21.6)), and similarly for $|X_n|$ and $|Y_n|$. ■

The random variables X_n are said to be *uniformly integrable* if

$$\lim_{\alpha \to \infty} \sup_n \int_{[|X_n| \ge \alpha]} |X_n| \, dP = 0; \tag{25.10}$$

see (16.21). This implies (see (16.22)) that

$$\sup_n E[|X_n|] < \infty. \tag{25.11}$$

Theorem 25.12. *If $X_n \Rightarrow X$ and the X_n are uniformly integrable, then X is integrable and*

$$E[X_n] \to E[X]. \tag{25.12}$$

PROOF. Construct random variables Y_n and Y as in the preceding proof. Since $Y_n \to Y$ with probability 1 and the Y_n are uniformly integrable in the sense of (16.21), $E[X_n] = E[Y_n] \to E[Y] = E[X]$ by Theorem 16.14. ■

If $\sup_n E[|X_n|^{1+\epsilon}] < \infty$ for some positive ϵ, then the X_n are uniformly integrable because

$$\int_{[|X_n| \ge \alpha]} |X_n| \, dP \le \frac{1}{\alpha^{\epsilon}} E\left[|X_n|^{1+\epsilon}\right]. \tag{25.13}$$

Since $X_n \Rightarrow X$ implies that $X_n^r \Rightarrow X^r$ by Theorem 25.7, there is the following consequence of the theorem.

Corollary. *Let r be a positive integer. If $X_n \Rightarrow X$ and $\sup_n E[|X_n|^{r+\epsilon}] < \infty$, where $\epsilon > 0$, then $E[|X|^r] < \infty$ and $E[X_n^r] \to E[X^r]$.*

The X_n are also uniformly integrable if there is an integrable random variable Z such that $P[|X_n| \ge t] \le P[|Z| \ge t]$ for $t > 0$, because then (21.10)

gives

$$\int_{[|X_n|\geq\alpha]} |X_n|\,dP \leq \int_{[|Z|\geq\alpha]} |Z|\,dP.$$

From this the dominated convergence theorem follows again.

PROBLEMS

25.1. (a) Show by example that distribution functions having densities can converge weakly even if the densities do not converge: *Hint:* Consider $f_n(x) = 1 + \cos 2\pi nx$ on $[0,1]$.

(b) Let f_n be 2^n times the indicator of the set of x in the unit interval for which $d_{n+1}(x) = \cdots = d_{2n}(x) = 0$, where $d_k(x)$ is the kth dyadic digit. Show that $f_n(x) \to 0$ except on a set of Lebesgue measure 0; on this exceptional set, redefine $f_n(x) = 0$ for all n, so that $f_n(x) \to 0$ everywhere. Show that the distributions corresponding to these densities converge weakly to Lebesgue measure confined to the unit interval.

(c) Show that distributions with densities can converge weakly to a limit that has no density (even to a unit mass).

(d) Show that discrete distributions can converge weakly to a distribution that has a density.

(e) Construct an example, like that of Example 25.3, in which $\mu_n(A) \to \mu(A)$ fails but in which all the measures come from continuous densities on $[0,1]$.

25.2. 14.8↑ Give a simple proof of the Gilvenko–Cantelli theorem (Theorem 20.6) under the extra hypothesis that F is continuous.

25.3. *Initial digits.* (a) Show that the first significant digit of a positive number x is d (in the scale of 10) if and only if $\{\log_{10} x\}$ lies between $\log_{10} d$ and $\log_{10}(d+1)$, $d = 1, \ldots, 9$, where the braces denote fractional part.

(b) For positive numbers $x_1, x_2, \ldots$, let $N_n(d)$ be the number among the first n that have initial digit d. Show that

$$\lim_n \frac{1}{n} N_n(d) = \log_{10}(d+1) - \log_{10} d, \qquad d = 1, \ldots, 9, \tag{25.14}$$

if the sequence $\log_{10} x_n$, $n = 1, 2, \ldots$, is uniformly distributed modulo 1. This is true, for example, of $x_n = \vartheta^n$ if $\log_{10} \vartheta$ is irrational.

(c) Let D_n be the first significant digit of a positive random variable X_n. Show that

$$\lim_n P[D_n = d] = \log_{10}(d+1) - \log_{10} d, \qquad d = 1, \ldots, 9, \tag{25.15}$$

if $\{\log_{10} X_n\} \Rightarrow U$, where U is uniformly distributed over the unit interval.

25.4. Show that for each probability measure μ on the line there exist probability measures μ_n with finite support such that $\mu_n \Rightarrow \mu$. Show further that $\mu_n\{x\}$ can

be taken rational and that each point in the support can be taken rational. Thus there exists a countable set of probability measures such that every μ is the weak limit of some sequence from the set. The space of distribution functions is thus separable in the Lévy metric (see Problem 14.5).

25.5. Show that (25.5) implies that $P([X \le x] \triangle [X_n \le x]) \to 0$ if $P[X = x] = 0$.

25.6. For arbitrary random variables X_n there exist positive constants a_n such that $a_n X_n \Rightarrow 0$.

25.7. Generalize Example 25.8 by showing for three-dimensional random vectors (A_n, B_n, X_n) and constants a and b, $a \ge 0$, that, if $A_n \Rightarrow a$, $B_n \Rightarrow b$, and $X_n \Rightarrow X$, then $A_n X_n + B_n \Rightarrow aX + b$. *Hint:* First show that if $Y_n \Rightarrow Y$ and $D_n \Rightarrow 0$, then $D_n Y_n \Rightarrow 0$.

25.8. Suppose that $X_n \Rightarrow X$ and that h_n and h are Borel functions. Let E be the set of x for which $h_n x_n \to hx$ fails for some sequence $x_n \to x$. Suppose that $E \in \mathscr{R}^1$ and $P[X \in E] = 0$. Show that $h_n X_n \Rightarrow hX$.

25.9. Suppose that the distributions of random variables X_n and X have densities f_n and f. Show that if $f_n(x) \to f(x)$ for x outside a set of Lebesgue measure 0, then $X_n \Rightarrow X$.

25.10. ↑ Suppose that X_n assumes as values $\gamma_n + k\delta_n$, $k = 0, \pm 1, \ldots$, where $\delta_n > 0$. Suppose that $\delta_n \to 0$ and that, if k_n is an integer varying with n in such a way that $\gamma_n + k_n\delta_n \to x$, then $P[X_n = \gamma_n + k_n\delta_n]\delta_n^{-1} \to f(x)$, where f is the density of a random variable X. Show that $X_n \Rightarrow X$.

25.11. ↑ Let S_n have the binomial distribution with parameters n and p. Assume as known that

$$P[S_n = k_n](np(1-p))^{1/2} \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{25.16}$$

if $(k_n - np)(np(1-p))^{-1/2} \to x$. Deduce the DeMoivre–Laplace theorem: $(S_n - np)(np(1-p))^{-1/2} \Rightarrow N$, where N has the standard normal distribution. This is a special case of the central limit theorem; see Section 27.

25.12. Prove weak convergence in Example 25.3 by using Theorem 25.8 and the theory of the Riemann integral.

25.13. (a) Show that probability measures satisfy $\mu_n \Rightarrow \mu$ if $\mu_n(a, b] \to \mu(a, b]$ whenever $\mu\{a\} = \mu\{b\} = 0$.

(b) Show that, if $\int f d\mu_n \to \int f d\mu$ for all continuous f with bounded support, then $\mu_n \Rightarrow \mu$.

25.14. ↑ Let μ be Lebesgue measure confined to the unit interval; let μ_n correspond to a mass of $x_{n,i} - x_{n,i-1}$ at some point in $(x_{n,i-1}, x_{n,i}]$, where $0 = x_{n0} < x_{n1} < \cdots < x_{nn} = 1$. Show by considering the distribution functions that $\mu_n \Rightarrow \mu$ if $\max_{i \le n}(x_{n,i} - x_{n,i-1}) \to 0$. Deduce that a bounded Borel function continuous almost everywhere on the unit interval is Riemann integrable. See Problem 17.1.

25.15. 2.18 5.19↑ A function f of positive integers has *distribution function* F if F is the weak limit of the distribution function $P_n[m: f(m) \le x]$ of f under the measure having probability $1/n$ at each of $1, \ldots, n$ (see 2.34)). In this case $D[m: f(m) \le x] = F(x)$ (see (2.35)) for continuity points x of F. Show that $\varphi(m)/m$ (see (2.37)) has a distribution:

(a) Show by the mapping theorem that it suffices to prove that $f(m) = \log(\varphi(m)/m) = \sum_p \delta_p(m) \log(1 - 1/p)$ has a distribution.

(b) Let $f_u(m) = \sum_{p \le u} \delta_p(m) \log(1 - 1/p)$, and show by (5.45) that f_u has distribution function $F_u(x) = P[\sum_{p \le u} X_p \log(1 - 1/p) \le x]$, where the X_p are independent random variables (one for each prime p) such that $P[X_p = 1] = 1/p$ and $P[X_p = 0] = 1 - 1/p$.

(c) Show that $\sum_p X_p \log(1 - 1/p)$ converges with probability 1. *Hint:* Use Theorem 22.6.

(d) Show that $\lim_{u \to \infty} \sup_n E_n[|f - f_u|] = 0$ (see (5.46) for the notation).

(e) Conclude by Markov's inequality and Theorem 25.5 that f has the distribution of the sum in (c).

25.16. For $A \in \mathscr{R}^1$ and $T > 0$, put $\lambda_T(A) = \lambda([-T, T] \cap A)/2T$, where λ is Lebesgue measure. The *relative measure* of A is

$$\rho(A) = \lim_{T \to \infty} \lambda_T(A), \tag{25.17}$$

provided that this limit exists. This is a continuous analogue of density (see (2.35)) for sets of integers. A Borel function f has a distribution under λ_T; if this converges weakly to F, then

$$\rho[x: f(x) \le u] = F(u) \tag{25.18}$$

for continuity points u of F, and F is called the distribution function of f. Show that all periodic functions have distributions.

25.17. Suppose that $\sup_n \int f \, d\mu_n < \infty$ for a nonnegative f such that $f(x) \to \infty$ as $x \to \pm\infty$. Show that $\{\mu_n\}$ is tight.

25.18. 23.4↑ Show that the random variables A_t and L_t in Problems 23.3 and 23.4 converge in distribution. Show that the moments converge.

25.19. In the applications of Theorem 9.2, only a weaker result is actually needed: For each K there exists a positive $\alpha = \alpha(K)$ such that if $E[X] = 0$, $E[X^2] = 1$, and $E[X^4] \le K$, then $P[X \ge 0] \ge \alpha$. Prove this by using tightness and the corollary to Theorem 25.12.

25.20. Find uniformly integrable random variables X_n for which there is no integrable Z satisfying $P[|X_n| \ge t] \le P[|Z| \ge t]$ for $t > 0$.

SECTION 26. CHARACTERISTIC FUNCTIONS

Definition

The *characteristic function* of a probability measure μ on the line is defined for real t by

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx}\mu(dx)$$

$$= \int_{-\infty}^{\infty} \cos tx\mu(dx) + i\int_{-\infty}^{\infty} \sin tx\mu(dx);$$

see the end of Section 16 for integrals of complex-valued functions.† A random variable X with distribution μ has characteristic function

$$\varphi(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx}\mu(dx).$$

The characteristic function is thus defined as the moment generating function but with the real argument s replaced by it; it has the advantage that it always exists because e^{itx} is bounded. The characteristic function in nonprobabilistic contexts is called the *Fourier transform*.

The characteristic function has three fundamental properties to be established here:

(i) If μ_1 and μ_2 have respective characteristic functions $\varphi_1(t)$ and $\varphi_2(t)$, then $\mu_1 * \mu_2$ has characteristic function $\varphi_1(t)\varphi_2(t)$. Although convolution is essential to the study of sums of independent random variables, it is a complicated operation, and it is often simpler to study the products of the corresponding characteristic functions.

(ii) The characteristic function uniquely determines the distribution. This shows that in studying the products in (i), no information is lost.

(iii) From the pointwise convergence of characteristic functions follows the weak convergence of the corresponding distributions. This makes it possible, for example, to investigate the asymptotic distributions of sums of independent random variables by means of their characteristic functions.

Moments and Derivatives

It is convenient first to study the relation between a characteristic function and the moments of the distribution it comes from.

†From complex variable theory only De Moivre's formula and the simplest properties of the exponential function are needed here.

Of course, $\varphi(0) = 1$, and by (16.30), $|\varphi(t)| \le 1$ for all t. By Theorem 16.8(i), $\varphi(t)$ is continuous in t. In fact, $|\varphi(t+h) - \varphi(t)| \le \int |e^{ihx} - 1| \mu(dx)$, and so it follows by the bounded convergence theorem that $\varphi(t)$ *is uniformly continuous*.

In the following relations, versions of Taylor's formula with remainder, x is assumed real. Integration by parts shows that

$$\int_0^x (x-s)^n e^{is}\, ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1}\int_0^x (x-s)^{n+1} e^{is}\, ds, \tag{26.1}$$

and it follows by induction that

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}\, ds \tag{26.2}$$

for $n \ge 0$. Replace n by $n-1$ in (26.1), solve for the integral on the right, and substitute this for the integral in (26.2); this gives

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is} - 1)\, ds. \tag{26.3}$$

Estimating the integrals in (26.2) and (26.3) (consider separately the cases $x \ge 0$ and $x < 0$) now leads to

$$\left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \le \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\} \tag{26.4}$$

for $n \ge 0$. The first term on the right gives a sharp estimate for $|x|$ small, the second a sharp estimate for $|x|$ large. For $n = 0, 1, 2$, the inequality specializes to

$$|e^{ix} - 1| \le \min\{|x|, 2\}, \tag{26.4$_0$}$$

$$|e^{ix} - (1 + ix)| \le \min\{\tfrac{1}{2}x^2, 2|x|\}, \tag{26.4$_1$}$$

$$|e^{ix} - (1 + ix - \tfrac{1}{2}x^2)| \le \min\{\tfrac{1}{6}|x|^3, x^2\}. \tag{26.4$_2$}$$

If X has a moment of order n, it follows that

$$\left| \varphi(t) - \sum_{k=0}^{n} \frac{(it)^k}{k!} E[X^k] \right| \le E\left[\min\left\{ \frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!} \right\} \right]. \tag{26.5}$$

For any t satisfying

$$\lim_n \frac{|t|^n E[|X|^n]}{n!} = 0, \tag{26.6}$$

$\varphi(t)$ must therefore have the expansion

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} E[X^k]; \tag{26.7}$$

compare (21.22). If

$$\sum_{k=0}^{\infty} \frac{|t|^k}{k!} E\left[|X|^k\right] = E[e^{|tX|}] < \infty,$$

then (see (16.31)) (26.7) must hold. Thus (26.7) holds if X has a moment generating function over the whole line.

Example 26.1. Since $E[e^{|tX|}] < \infty$ if X has the standard normal distribution, by (26.7) and (21.7) its characteristic function is

(26.8)

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(it)^{2k}}{(2k)!} 1 \times 3 \times \cdots \times (2k-1) = \sum_{k=0}^{\infty} \frac{1}{k!} \left(-\frac{t^2}{2}\right)^k = e^{-t^2/2}.$$

This and (21.25) formally coincide if $s = it$. ■

If the power-series expansion (26.7) holds, the moments of X can be read off from it:

$$\varphi^{(k)}(0) = i^k E[X^k]. \tag{26.9}$$

This is the analogue of (21.23). It holds, however, under the weakest possible assumption, namely that $E[|X^k|] < \infty$. Indeed,

$$\frac{\varphi(t+h) - \varphi(t)}{h} - E[iXe^{itX}] = E\left[e^{itX} \frac{e^{ihX} - 1 - ihX}{h}\right].$$

By (26.4_1), the integrand on the right is dominated by $2|X|$ and goes to 0 with h; hence the expected value goes to 0 by the dominated convergence theorem. Thus $\varphi'(t) = E[iXe^{itX}]$. Repeating this argument inductively gives

$$\varphi^{(k)}(t) = E\left[(iX)^k e^{itX}\right] \tag{26.10}$$

if $E[|X^k|] < \infty$. Hence (26.9) holds if $E[|X^k|] < \infty$. The proof of uniform continuity for $\varphi(t)$ works for $\varphi^{(k)}(t)$ as well.

If $E[X^2]$ is finite, then

$$\varphi(t) = 1 + itE[X] - \tfrac{1}{2}t^2E[X^2] + o(t^2), \qquad t \to 0. \tag{26.11}$$

Indeed, by (26.4_2), the error is at most $t^2E[\min\{|t||X|^3, X^2\}]$, and as $t \to 0$ the integrand goes to 0 and is dominated by X^2. Estimates of this kind are essential for proving limit theorems.

The more moments μ has, the more derivatives φ has. This is one sense in which lightness of the tails of μ is reflected by smoothness of φ. There are results which connect the behavior of $\varphi(t)$ as $|t| \to \infty$ with smoothness properties of μ. The *Riemann–Lebesgue theorem* is the most important of these:

Theorem 26.1. *If μ has a density, then $\varphi(t) \to 0$ as $|t| \to \infty$.*

PROOF. The problem is to prove for integrable f that $\int f(x)e^{itx}\,dx \to 0$ as $|t| \to \infty$. There exists by Theorem 17.1 a step function $g = \sum_k \alpha_k I_{A_k}$, a finite linear combination of indicators of intervals $A_k = (a_k, b_k]$, for which $\int |f - g|\,dx < \epsilon$. Now $\int f(x)e^{itx}\,dx$ differs by at most ϵ from $\int g(x)e^{itx}\,dx = \sum_k \alpha_k(e^{itb_k} - e^{ita_k})/it$, and this goes to 0 as $|t| \to \infty$. ■

Independence

The multiplicative property (21.28) of moment generating functions extends to characteristic functions. Suppose that X_1 and X_2 are independent random variables with characteristic functions φ_1 and φ_2. If $Y_j = \cos X_j$ and $Z_j = \sin tX_j$, then (Y_1, Z_1) and (Y_2, Z_2) are independent; by the rules for integrating complex-valued functions,

$$\begin{aligned}\varphi_1(t)\varphi_2(t) &= (E[Y_1] + iE[Z_1])(E[Y_2] + iE[Z_2]) \\ &= E[Y_1]E[Y_2] - E[Z_1]E[Z_2] \\ &\quad + i(E[Y_1]E[Z_2] + E[Z_1]E[Y_2]) \\ &= E[Y_1Y_2 - Z_1Z_2 + i(Y_1Z_2 + Z_1Y_2)] = E[e^{it(X_1+X_2)}].\end{aligned}$$

This extends to sums of three or more: If $X_1, \ldots, X_n$ are independent, then

$$E[e^{it\sum_{k=1}^n X_k}] = \prod_{k=1}^n E[e^{itX_k}]. \tag{26.12}$$

If X has characteristic function $\varphi(t)$, then $aX+b$ has characteristic function

$$E\left[e^{it(aX+b)}\right] = e^{itb}\varphi(at). \tag{26.13}$$

In particular, $-X$ has characteristic function $\varphi(-t)$, which is the complex conjugate of $\varphi(t)$.

Inversion and the Uniqueness Theorem

A characteristic function φ uniquely determines the measure μ it comes from. This fundamental fact will be derived by means of an inversion formula through which μ can in principle be recovered from φ.

Define

$$S(T) = \int_0^T \frac{\sin x}{x}\, dx, \qquad T \geq 0.$$

In Example 18.4 it is shown that

$$\lim_{T\to\infty} S(T) = \frac{\pi}{2}; \tag{26.14}$$

$S(T)$ is therefore bounded. If $\operatorname{sgn}\theta$ is $+1$, 0, or -1 as θ is positive, 0, or negative, then

$$\int_0^T \frac{\sin t\theta}{t}\, dt = \operatorname{sgn}\theta \cdot S(T|\theta|), \qquad T \geq 0. \tag{26.15}$$

Theorem 26.2. *If the probability measure μ has characteristic function φ, and if $\mu\{a\} = \mu\{b\} = 0$, then*

$$\mu(a,b] = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt. \tag{26.16}$$

Distinct measures cannot have the same characteristic function.

Note: By (26.4_1) the integrand here converges as $t \to 0$ to $b-a$, which is to be taken as its value for $t=0$. For fixed a and b the integrand is thus continuous in t, and by (26.4_0) it is bounded. If μ is a unit mass at 0, then $\varphi(t) \equiv 1$ and the integral in (26.16) cannot be extended over the whole line.

PROOF. The inversion formula will imply uniqueness: It will imply that if μ and ν have the same characteristic function, then $\mu(a,b] = \nu(a,b]$ if $\mu\{a\} = \nu\{a\} = \mu\{b\} = \nu\{b\} = 0$; but such intervals $(a,b]$ form a π-system generating $\mathscr{R}^1$.

Denote by I_T the quantity inside the limit in (26.16). By Fubini's theorem

$$I_T = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\int_{-T}^{T}\frac{e^{it(x-a)}-e^{it(x-b)}}{it}\,dt\right]\mu(dx). \tag{26.17}$$

This interchange is legitimate because the double integral extends over a set of finite product measure and by (26.4_0) the integrand is bounded by $|b-a|$. Rewrite the integrand by DeMoivre's formula. Since $\sin s$ and $\cos s$ are odd and even, respectively, (26.15) gives

$$I_T = \int_{-\infty}^{\infty}\left[\frac{\operatorname{sgn}(x-a)}{\pi}S(T\cdot|x-a|) - \frac{\operatorname{sgn}(x-b)}{\pi}S(T\cdot|x-b|)\right]\mu(dx).$$

The integrand here is bounded and converges as $T\to\infty$ to the function

$$\psi_{a,b}(x) = \begin{cases} 0 & \text{if } x<a,\\ \frac{1}{2} & \text{if } x=a,\\ 1 & \text{if } a<x<b,\\ \frac{1}{2} & \text{if } x=b,\\ 0 & \text{if } b<x. \end{cases} \tag{26.18}$$

Thus $I_T\to\int\psi_{a,b}\,d\mu$, which implies that (26.16) holds if $\mu\{a\}=\mu\{b\}=0$. ■

The inversion formula contains further information. Suppose that

$$\int_{-\infty}^{\infty}|\varphi(t)|\,dt<\infty. \tag{26.19}$$

In this case the integral in (26.16) can be extended over R^1. By (26.4_0),

$$\left|\frac{e^{-itb}-e^{-ita}}{it}\right| = \frac{|e^{it(b-a)}-1|}{|t|} \le |b-a|;$$

therefore, $\mu(a,b)\le(b-a)\int_{-\infty}^{\infty}|\varphi(t)|\,dt$, and there can be no point masses. By (26.16), the corresponding distribution function satisfies

$$\frac{F(x+h)-F(x)}{h} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-itx}-e^{-it(x+h)}}{ith}\varphi(t)\,dt$$

(whether h is positive or negative). The integrand is by (26.4_0) dominated by $|\varphi(t)|$ and goes to $e^{itx}\varphi(t)$ as $h\to 0$. Therefore, F has derivative

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}\varphi(t)\,dt. \tag{26.20}$$

Since f is continuous for the same reason φ is, it integrates to F by the fundamental theorem of the calculus (see (17.6)). Thus (26.19) *implies that* μ *has the continuous density* (26.20). Moreover, this is the only continuous density. In this result, as in the Riemann–Lebesgue theorem, conditions on the size of $\varphi(t)$ for large $|t|$ are connected with smoothness properties of μ.

The inversion formula (26.20) has many applications. In the first place, it can be used for a new derivation of (26.14). As pointed out in Example 17.3, the existence of the limit in (26.14) is easy to prove. Denote this limit temporarily by $\pi_0/2$—without assuming that $\pi_0 = \pi$. Then (26.16) and (26.20) follow as before if π is replaced by π_0. Applying the latter to the standard normal density (see (26.8)) gives

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \frac{1}{2\pi_0} \int_{-\infty}^{\infty} e^{-itx} e^{-t^2/2}\, dt, \tag{26.21}$$

where the π on the left is that of analysis and geometry—it comes ultimately from the quadrature (18.10). An application of (26.8) with x and t interchanged reduces the right side of (26.21) to $(\sqrt{2\pi}/2\pi_0)e^{-x^2/2}$, and therefore π_0 does equal π.

Consider the densities in the table. The characteristic function for the *normal* distribution has already been calculated. For the *uniform* distribution over $(0, 1)$, the computation is of course straightforward; note that in this case the density cannot be recovered from (26.20), because $\varphi(t)$ is not integrable; this is reflected in the fact that the density has discontinuities at 0 and 1.

Distribution	Density	Interval	Characteristic Function
1. Normal	$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$-\infty < x < \infty$	$e^{-t^2/2}$
2. Uniform	1	$0 < x < 1$	$\frac{e^{it} - 1}{it}$
3. Exponential	e^{-x}	$0 < x < \infty$	$\frac{1}{1 - it}$
4. Double exponential or Laplace	$\frac{1}{2} e^{-\lvert x \rvert}$	$-\infty < x < \infty$	$\frac{1}{1 + t^2}$
5. Cauchy	$\frac{1}{\pi} \frac{1}{1 + x^2}$	$-\infty < x < \infty$	$e^{-\lvert t \rvert}$
6. Triangular	$1 - \lvert x \rvert$	$-1 < x < 1$	$2\frac{1 - \cos t}{t^2}$
7.	$\frac{1}{\pi} \frac{1 - \cos x}{x^2}$	$-\infty < x < \infty$	$(1 - \lvert t \rvert) I_{(-1,1)}(t)$

The characteristic function for the *exponential* distribution is easily calculated; compare Example 21.3. As for the *double exponential* or *Laplace* distribution, $e^{-|x|}e^{itx}$ integrates over $(0,\infty)$ to $(1-it)^{-1}$ and over $(-\infty,0)$ to $(1+it)^{-1}$, which gives the result. By (26.20), then,

$$e^{-|x|} = \frac{1}{\pi}\int_{-\infty}^{\infty} e^{-itx}\frac{dt}{1+t^2}.$$

For $x=0$ this gives the standard integral $\int_{-\infty}^{\infty} dt/(1+t^2) = \pi$; see Example 17.5. Thus the *Cauchy* density in the table integrates to 1 and has characteristic function $e^{-|t|}$. This distribution has no first moment, and the characteristic function is not differentiable at the origin.

A straightforward integration shows that the *triangular* density has the characteristic function given in the table, and by (26.20),

$$(1-|x|)I_{(-1,1)}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} e^{-itx}\frac{1-\cos t}{t^2}\,dt.$$

For $x=0$ this is $\int_{-\infty}^{\infty}(1-\cos t)t^{-2}\,dt = \pi$; hence the last line of the table.

Each density and characteristic function in the table can be transformed by (26.13), which gives a family of distributions.

The Continuity Theorem

Because of (26.12), the characteristic function provides a powerful means of studying the distributions of sums of independent random variables. It is often easier to work with products of characteristic functions than with convolutions, and knowing the characteristic function of the sum is by Theorem 26.2 in principle the same thing as knowing the distribution itself. Because of the following *continuity theorem*, characteristic functions can be used to study limit distributions.

Theorem 26.3. *Let μ_n, μ be probability measures with characteristic functions φ_n, φ. A necessary and sufficient condition for $\mu_n \Rightarrow \mu$ is that $\varphi_n(t) \to \varphi(t)$ for each t.*

PROOF. *Necessity*. For each t, e^{itx} has bounded modulus and is continuous in x. The necessity therefore follows by an application of Theorem 25.8 (to the real and imaginary parts of e^{itx}).

Sufficiency. By Fubini's theorem,

$$(26.22)\qquad \frac{1}{u}\int_{-u}^{u}(1-\varphi_n(t))\,dt = \int_{-\infty}^{\infty}\left[\frac{1}{u}\int_{-u}^{u}(1-e^{itx})\,dt\right]\mu_n(dx)$$

$$= 2\int_{-\infty}^{\infty}\left(1-\frac{\sin ux}{ux}\right)\mu_n(dx)$$

$$\geq 2\int_{|x|\geq 2/u}\left(1-\frac{1}{|ux|}\right)\mu_n(dx)$$

$$\geq \mu_n\left[x\colon |x|\geq \frac{2}{u}\right].$$

(Note that the first integral is real.) Since φ is continuous at the origin and $\varphi(0)=1$, there is for positive ϵ a u for which $u^{-1}\int_{-u}^{u}(1-\varphi(t))\,dt<\epsilon$. Since φ_n converges to φ, the bounded convergence theorem implies that there exists an n_0 such that $u^{-1}\int_{-u}^{u}(1-\varphi_n(t))\,dt<2\epsilon$ for $n\geq n_0$. If $a=2/u$ in (26.22), then $\mu_n[x\colon |x|\geq a]<2\epsilon$ for $n\geq n_0$. Increasing a if necessary will ensure that this inequality also holds for the finitely many n preceding n_0. Therefore, $\{\mu_n\}$ is tight.

By the corollary to Theorem 25.10, $\mu_n\Rightarrow\mu$ will follow if it is shown that each subsequence $\{\mu_{n_k}\}$ that converges weakly at all converges weakly to μ. But if $\mu_{n_k}\Rightarrow\nu$ as $k\to\infty$, then by the necessity half of the theorem, already proved, ν has characteristic function $\lim_k \varphi_{n_k}(t)=\varphi(t)$. By Theorem 26.2, ν and μ must coincide. ■

Two corollaries, interesting in themselves, will make clearer the structure of the proof of sufficiency given above. In each, let μ_n be probability measures on the line with characteristic functions φ_n.

Corollary 1. *Suppose that* $\lim_n \varphi_n(t)=g(t)$ *for each* t, *where the limit function* g *is continuous at* 0. *Then there exists a* μ *such that* $\mu_n\Rightarrow\mu$, *and* μ *has characteristic function* g.

PROOF. The point of the corollary is that g is not assumed at the outset to be a characteristic function. But in the argument following (26.22), only $\varphi(0)=1$ and the continuity of φ at 0 were used; hence $\{\mu_n\}$ is tight under the present hypothesis. If $\mu_{n_k}\Rightarrow\nu$ as $k\to\infty$, then ν must have characteristic function $\lim_k \varphi_{n_k}(t)=g(t)$. Thus g is, in fact, a characteristic function, and the proof goes through as before. ■

In this proof the continuity of g was used to establish tightness. Hence if $\{\mu_n\}$ is assumed tight in the first place, the hypothesis of continuity can be suppressed:

Corollary 2. *Suppose that* $\lim_n \varphi_n(t) = g(t)$ *exists for each* t *and that* $\{\mu_n\}$ *is tight. Then there exists a* μ *such that* $\mu_n \Rightarrow \mu$, *and* μ *has characteristic function* g.

This second corollary applies, for example, if the μ_n have a common bounded support.

Example 26.2. If μ_n is the uniform distribution over $(-n, n)$, its characteristic function is $(nt)^{-1}\sin tn$ for $t \neq 0$, and hence it converges to $I_{\{0\}}(t)$. In this case $\{\mu_n\}$ is not tight, the limit function is not continuous at 0, and μ_n does not converge weakly. ■

Fourier Series*

Let μ be a probability measure on $\mathscr{R}^1$ that is supported by $[0, 2\pi]$. Its Fourier coefficients are defined by

$$c_m = \int_0^{2\pi} e^{imx}\mu(dx), \qquad m = 0, \pm 1, \pm 2, \ldots . \tag{26.23}$$

These coefficients, the values of the characteristic function for integer arguments, suffice to determine μ except for the weights it may put at 0 and 2π. The relation between μ and its Fourier coefficients can be expressed formally by

$$\mu(dx) \sim \frac{1}{2\pi}\sum_{l=-\infty}^{\infty} c_l e^{-ilx}\,dx: \tag{26.24}$$

if the $\mu(dx)$ in (26.23) is replaced by the right side of (26.24), and if the sum over l is interchanged with the integral, the result is a formal identity.

To see how to recover μ from its Fourier coefficients, consider the symmetric partial sums $s_m(t) = (2\pi)^{-1}\sum_{l=-m}^{m} c_l e^{-ilt}$ and their Cesàro averages $\sigma_m(t) = m^{-1}\sum_{l=0}^{m-1} s_l(t)$. From the trigonometric identity [A24]

$$\sum_{l=0}^{m-1}\sum_{k=-1}^{l} e^{ikx} = \frac{\sin^2 \frac{1}{2}mx}{\sin^2 \frac{1}{2}x} \tag{26.25}$$

it follows that

$$\sigma_m(t) = \frac{1}{2\pi m}\int_0^{2\pi} \frac{\sin^2 \frac{1}{2}m(x-t)}{\sin^2 \frac{1}{2}(x-t)}\mu(dx). \tag{26.26}$$

*This topic may be omitted.

If μ is $(2\pi)^{-1}$ times Lebesgue measure confined to $[0, 2\pi]$, then $c_0 = 1$ and $c_m = 0$ for $m \neq 0$, so that $\sigma_m(t) = s_m(t) = (2\pi)^{-1}$; this gives the identity

$$\frac{1}{2\pi m}\int_{-\pi}^{\pi} \frac{\sin^2 \frac{1}{2}ms}{\sin^2 \frac{1}{2}s}\,ds = 1. \tag{26.27}$$

Suppose that $0 < a < b < 2\pi$, and integrate (26.26) over (a, b). Fubini's theorem (the integrand is nonnegative) and a change of variable lead to

$$\int_a^b \sigma_m(t)\,dt = \int_0^{2\pi}\left[\frac{1}{2\pi m}\int_{a-x}^{b-x} \frac{x \sin^2 \frac{1}{2}ms}{\sin^2 \frac{1}{2}s}\,ds\right]\mu(dx). \tag{26.28}$$

The denominator in (26.27) is bounded away from 0 outside $(-\delta, \delta)$, and so as m goes to ∞ with δ fixed $(0 < \delta < \pi)$,

$$\frac{1}{2\pi m}\int_{\delta<|s|<\pi} \frac{\sin^2 \frac{1}{2}ms}{\sin^2 \frac{1}{2}s}\,ds \to 0, \qquad \frac{1}{2\pi m}\int_{|s|<\delta} \frac{\sin^2 \frac{1}{2}ms}{\sin^2 \frac{1}{2}s}\,ds \to 1.$$

Therefore, the expression in brackets in (26.28) goes to 0 if $0 \le x < a$ or $b < x \le 2\pi$, and it goes to 1 if $a < x < b$; and because of (26.27), it is bounded by 1. It follows by the bounded convergence theorem that

$$\mu(a, b] = \lim_m \int_a^b \sigma_m(t)\,dt \tag{26.29}$$

if $\mu\{a\} = \mu\{b\} = 0$ and $0 < a < b < 2\pi$.

This is the analogue of (26.16). If μ and ν have the same Fourier coefficients, it follows from (26.29) that $\mu(A) = \nu(A)$ for $A \subset (0, 2\pi)$ and hence that $\mu\{0, 2\pi\} = \nu\{0, 2\pi\}$. It is clear from periodicity that the coefficients (26.23) are unchanged if $\mu\{0\}$ and $\mu\{2\pi\}$ are altered but $\mu\{0\} + \mu\{2\pi\}$ is held constant.

Suppose that μ_n is supported by $[0, 2\pi]$ and has coefficients $c_m^{(n)}$, and suppose that $\lim_n c_m^{(n)} = c_m$ for all m. Since $\{\mu_n\}$ is tight, $\mu_n \Rightarrow \mu$ will hold if $\mu_{n_k} \Rightarrow \nu$ $(k \to \infty)$ implies $\nu = \mu$. But in this case ν and μ have the same coefficients c_m, and hence they are identical except perhaps in the way they split the mass $\nu\{0, 2\pi\} = \mu\{0, 2\pi\}$ between the points 0 and 2π. But this poses no problem if $\mu\{0, 2\pi\} = 0$: *If* $\lim_n c_m^{(n)} = c_m$ *for all* m *and* $\mu\{0\} = \mu\{2\pi\} = 0$, *then* $\mu_n \Rightarrow \mu$.

Example 26.3. If μ is $(2\pi)^{-1}$ times Lebesgue measure confined to the interval $[0, 2\pi]$, the condition is that $\lim_n c_m^{(n)} = 0$ for $m \neq 0$. Let $x_1, x_2, \ldots$ be a sequence of reals, and let μ_n put mass n^{-1} at each point $2\pi\{x_k\}$, $1 \le k \le n$, where $\{x_k\} = x_k - \lfloor x_k \rfloor$ denotes fractional part. This is the probability measure (25.3) rescaled to $[0, 2\pi]$. The sequence $x_1, x_2, \ldots$ is uniformly distributed modulo 1 if and only if

$$\frac{1}{n}\sum_{k=1}^{n} e^{2\pi i\{x_k\}m} = \frac{1}{n}\sum_{k=1}^{n} e^{2\pi i x_k m} \to 0$$

for $m \neq 0$. This is *Weyl's criterion*.

If $x_k = k\theta$, where θ is irrational, then $\exp(2\pi i\theta m) \neq 1$ for $m \neq 0$ and hence

$$\frac{1}{n}\sum_{k=1}^{n} e^{2\pi ik\theta m} = \frac{1}{n} e^{2\pi i\theta m}\frac{1 - e^{2\pi in\theta m}}{1 - e^{2\pi i\theta m}} \to 0.$$

Thus $\theta, 2\theta, 3\theta, \ldots$ is uniformly distributed modulo 1 if θ is irrational, which gives another proof of Theorem 25.1. ■

PROBLEMS

26.1. A random variable has a *lattice distribution* if for some a and b, $b > 0$, the lattice $[a + nb\colon n = 0, \pm 1, \ldots]$ supports the distribution of X. Let X have characteristic function φ.

(a) Show that a necessary condition for X to have a lattice distribution is that $|\varphi(t)| = 1$ for some $t \neq 0$.

(b) Show that the condition is sufficient as well.

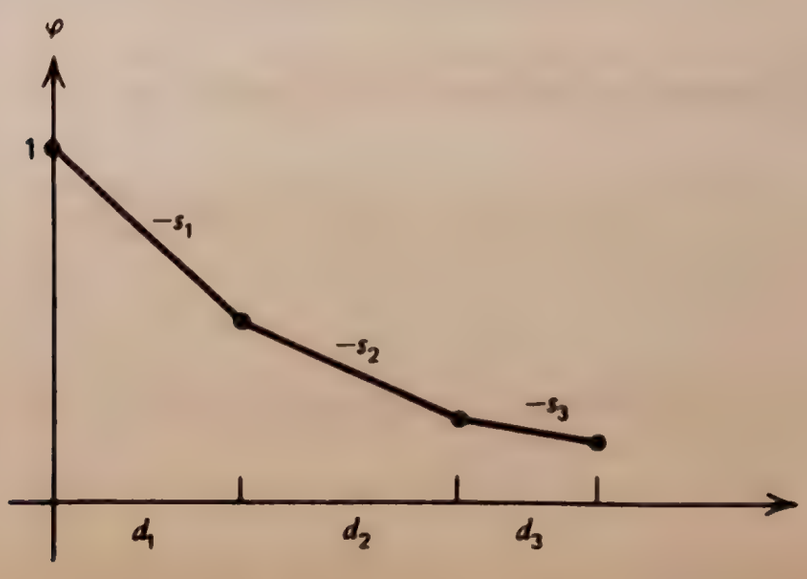
(c) Suppose that $|\varphi(t)| = |\varphi(t')| = 1$ for incommensurable t and t' ($t \neq 0$, $t' \neq 0$, t/t' irrational). Show that $P[X = c] = 1$ for some constant c.

26.2. If $\mu(-\infty, x] = \mu[-x, \infty)$ for all x (which implies that $\mu(A) = \mu(-A)$ for all $A \in \mathscr{R}^1$), then μ is *symmetric*. Show that this holds if and only if the characteristic function is real.

26.3. Consider functions φ that are real and nonnegative and satisfy $\varphi(-t) = \varphi(t)$ and $\varphi(0) = 1$.

(a) Suppose that $d_1, d_2, \ldots$ are positive and $\sum_{k=1}^{\infty} d_k = \infty$, that $s_1 \geq s_2 \geq \cdots \geq 0$ and $\lim_k s_k = 0$, and that $\sum_{k=1}^{\infty} s_k d_k = 1$. Let φ be the convex polygon whose successive sides have slopes $-s_1, -s_2, \ldots$ and lengths $d_1, d_2, \ldots$ when projected on the horizontal axis: φ has value $1 - \sum_{j=1}^{k} s_j d_j$ at $t_k = d_1 + \cdots + d_k$. If $s_n = 0$, there are in effect only n sides. Let $\varphi_0(t) = (1 - |t|)I_{(-1,1)}(t)$ be the characteristic function in the last line in the table on p. 348, and show that $\varphi(t)$ is a convex combination of the characteristic functions $\varphi_0(t/t_k)$ and hence is itself a characteristic function.

(b) *Pólya's criterion*. Show that φ is a characteristic function if it is even and continuous and, on $[0, \infty)$, nonincreasing and convex ($\varphi(0) = 1$).



26.4. ↑ Let φ_1 and φ_2 be characteristic functions, and show that the set $A = [t: \varphi_1(t) = \varphi_2(t)]$ is closed, contains 0, and is symmetric about 0. Show that every set with these three properties can be such an A. What does this say about the uniqueness theorem?

26.5. Show by Theorem 26.1 and integration by parts that if μ has a density f with integrable derivative f', then $\varphi(t) = o(t^{-1})$ as $|t| \to \infty$. Extend to higher derivatives.

26.6. Show for independent random variables uniformly distributed over $(-1, +1)$ that $X_1 + \cdots + X_n$ has density $\pi^{-1}\int_0^\infty ((\sin t)/t)^n \cos tx\, dt$ for $n \geq 2$.

26.7. 21.17↑ *Uniqueness theorem for moment generating functions.* Suppose that F has a moment generating function in $(-s_0, s_0)$, $s_0 > 0$. From the fact that $\int_{-\infty}^{\infty} e^{zx}\, dF(x)$ is analytic in the strip $-s_0 < \operatorname{Re} z < s_0$, prove that the moment generating function determines F. Show that it is enough that the moment generating function exist in $[0, s_0)$, $s_0 > 0$.

26.8. 21.20 26.7↑ Show that the gamma density (20.47) has characteristic function

$$\frac{1}{(1 - it/\alpha)^u} = \exp\left[-u \log\left(1 - \frac{it}{\alpha}\right)\right],$$

where the logarithm is the principal part. Show that $\int_0^\infty e^{zx} f(x; \alpha, u)\, dx$ is analytic for $\operatorname{Re} z < \alpha$.

26.9. Use characteristic functions for a simple proof that the family of Cauchy distributions defined by (20.45) is closed under convolution; compare the argument in Problem 20.14(a). Do the same for the normal distribution (compare Example 20.6) and for the Poisson and gamma distributions.

26.10. Suppose that $F_n \Rightarrow F$ and that the characteristic functions are dominated by an integrable function. Show that F has a density that is the limit of the densities of the F_n.

26.11. Show for all a and b that the right side of (26.16) is $\mu(a, b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\}$.

26.12. By the kind of argument leading to (26.16), show that

$$\mu\{a\} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita} \varphi(t)\, dt. \tag{26.30}$$

26.13. ↑ Let $x_1, x_2, \ldots$ be the points of positive μ-measure. By the following steps prove that

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |\varphi(t)|^2\, dt = \sum_k (\mu\{x_k\})^2. \tag{26.31}$$

Let X and Y be independent and have characteristic function φ.

(a) Show by (26.30) that the left side of (26.31) is $P[X - Y = 0]$.

(b) Show (Theorem 20.3) that $P[X - Y = 0] = \int_{-\infty}^{\infty} P[X = y]\mu(dy) = \sum_k (\mu\{x_k\})^2$.

26.14. ↑ Show that μ has no point masses if $\varphi^2(t)$ is integrable.

26.15. (a) Show that if $\{\mu_n\}$ is tight, then the characteristic functions $\varphi_n(t)$ are uniformly equicontinuous (for each ϵ there is a δ such that $|s - t| < \delta$ implies that $|\varphi_n(s) - \varphi_n(t)| < \epsilon$ for all n).

(b) Show that $\mu_n \Rightarrow \mu$ implies that $\varphi_n(t) \to \varphi(t)$ uniformly on bounded sets.

(c) Show that the convergence in part (b) need not be uniform over the entire line.

26.16. 14.5 26.15↑ For distribution functions F and G, define $d'(F, G) = \sup_t |\varphi(t) - \psi(t)|/(1 + |t|)$, where φ and ψ are the corresponding characteristic functions. Show that this is a metric and equivalent to the Lévy metric.

26.17. 25.16↑ A real function f has *mean value*

$$M[f(x)] = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(x)\, dx, \tag{26.32}$$

provided that f is integrable over each $[-T, T]$ and the limit exists.

(a) Show that, if f is bounded and $e^{itf(x)}$ has a mean value for each t, then f has a distribution in the sense of (25.18).

(b) Show that

$$M[e^{itx}] = \begin{cases} 1 & \text{if } t = 0, \\ 0 & \text{if } t \neq 0. \end{cases} \tag{26.33}$$

Of course, $f(x) = x$ has no distribution.

26.18. Suppose that X is irrational with probability 1. Let μ_n be the distribution of the fractional part $\{nX\}$. Use the continuity theorem and Theorem 25.1 to show that $n^{-1}\sum_{k=1}^{n} \mu_k$ converges weakly to the uniform distribution on $[0, 1]$.

26.19. 25.13↑ The uniqueness theorem for characteristic functions can be derived from the Weierstrass approximation theorem. Fill in the details of the following argument. Let μ and ν be probability measures on the line. For continuous f with bounded support choose a so that $\mu(-a, a)$ and $\nu(-a, a)$ are nearly 1 and f vanishes outside $(-a, a)$. Let g be periodic and agree with f in $(-a, a)$, and by the Weierstrass theorem uniformly approximate $g(x)$ by a trigonometric sum $p(x) = \sum_{k=1}^{N} a_k e^{it_k x}$. If μ and ν have the same characteristic function, then $\int f\, d\mu \approx \int g\, d\mu \approx \int p\, d\mu = \int p\, d\nu \approx \int g\, d\nu \approx \int f\, d\nu$.

26.20. Use the continuity theorem to prove the result in Example 25.2 concerning the convergence of the binomial distribution to the Poisson.

26.21. According to Example 25.8, if $X_n \Rightarrow X$, $a_n \to a$, and $b_n \to b$, then $a_n X_n + b_n \Rightarrow aX + b$. Prove this by means of characteristic functions.

26.22. 26.1 26.15↑ According to Theorem 14.2, if $X_n \Rightarrow X$ and $a_n X_n + b_n \Rightarrow Y$, where $a_n > 0$ and the distributions of X and Y are nondegenerate, then $a_n \to a > 0$, $b_n \to b$, and $aX + b$ and Y have the same distribution. Prove this by characteristic functions. Let φ_n, φ, ψ be the characteristic functions of X_n, X, Y.

(a) Show that $|\varphi_n(a_n t)| \to |\psi(t)|$ uniformly on bounded sets and hence that a_n cannot converge to 0 along a subsequence.

(b) Interchange the roles of φ and ψ and show that a_n cannot converge to infinity along a subsequence.

(c) Show that a_n converges to some $a > 0$.

(d) Show that $e^{itb_n} \to \psi(t)/\varphi(at)$ in a neighborhood of 0 and hence that $\int_0^t e^{isb_n}\, ds \to \int_0^t (\psi(s)/\varphi(as))\, ds$. Conclude that b_n converges.

26.23. Prove a continuity theorem for moment generating functions as defined by (22.4) for probability measures on $[0, \infty)$. For uniqueness, see Theorem 22.2; the analogue of (26.22) is

$$\frac{2}{u}\int_0^u (1 - M(s))\, ds \geq \mu\left(\frac{2}{u}, \infty\right).$$

26.24. 26.4↑ Show by example that the values $\varphi(m)$ of the characteristic function at integer arguments may not determine the distribution if it is not supported by $[0, 2\pi]$.

26.25. If f is integrable over $[0, 2\pi]$, define its Fourier coefficients as $\int_0^{2\pi} e^{imx} f(x)\, dx$. Show that these coefficients uniquely determine f up to sets of measure 0.

26.26. 19.8 26.25↑ Show that the trigonometric system (19.17) is complete.

26.27. The Fourier-series analogue of the condition (26.19) is $\sum_m |c_m| < \infty$. Show that it implies μ has density $f(x) = (2\pi)^{-1} \sum_m c_m e^{-imx}$ on $[0, 2\pi]$, where f is continuous and $f(0) = f(2\pi)$. This is the analogue of the inversion formula (26.20).

26.28. ↑ Show that

$$(\pi - x)^2 = \frac{\pi^2}{3} + 4\sum_{m=1}^{\infty} \frac{\cos mx}{m^2}, \qquad 0 \leq x \leq 2\pi.$$

Show that $\sum_{m=1}^{\infty} 1/m^2 = \pi^2/6$ and $\sum_{m=1}^{\infty} (-1)^{m+1}/m^2 = \pi^2/12$.

26.29. (a) Suppose X' and X'' are independent random variables with values in $[0, 2\pi]$, and let X be $X' + X''$ reduced module 2π. Show that the corresponding Fourier coefficients satisfy $c_m = c'_m c''_m$.

(b) Show that if one or the other of X' and X'' is uniformly distributed, so is X.

26.30. 26.25↑ The theory of Fourier series can be carried over from $[0, 2\pi]$ to the unit circle in the complex plane with normalized circular Lebesgue measure P. The circular functions e^{imx} become the powers ω^m, and an integrable f is determined to within sets of measure 0 by its Fourier coefficients $c_m = \int_\Omega \omega^m f(\omega) P(d\omega)$. Suppose that A is invariant under the rotation through the angle $\arg c$ (Example 24.4). Find a relation on the Fourier coefficients of I_A, and conclude that the rotation is ergodic if c is not a root of unity. Compare the proof on p. 316.

SECTION 27. THE CENTRAL LIMIT THEOREM

Identically Distributed Summands

The central limit theorem says roughly that the sum of many independent random variables will be approximately normally distributed if each summand has high probability of being small. Theorem 27.1, the *Lindeberg–Lévy theorem*, will give an idea of the techniques and hypotheses needed for the more general results that follow.

Throughout, N will denote a random variable with the standard normal distribution:

$$P[N \in A] = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2}\, dx. \tag{27.1}$$

Theorem 27.1. *Suppose that $\{X_n\}$ is an independent sequence of random variables having the same distribution with mean c and finite positive variance σ^2. If $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n - nc}{\sigma\sqrt{n}} \Rightarrow N. \tag{27.2}$$

By the argument in Example 25.7, (27.2) implies that $n^{-1}S_n \Rightarrow c$. The central limit theorem and the strong law of large numbers thus refine the weak law of large numbers in different directions.

Since Theorem 27.1 is a special case of Theorem 27.2, no proof is really necessary. To understand the methods of this section, however, consider the special case in which X_k takes the values ± 1 with probability $1/2$ each. Each X_k then has characteristic function $\varphi(t) = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos t$. By (26.12) and (26.13), $S_n/\sqrt{n}$ has characteristic function $\varphi^n(t/\sqrt{n})$, and so, by the continuity theorem, the problem is to show that $\cos^n t/\sqrt{n} \to E[e^{itN}] = e^{-t^2/2}$, or that $n \log\cos t/\sqrt{n}$ (well defined for large n) goes to $-\frac{1}{2}t^2$. But this follows by l'Hopital's rule: Let $t/\sqrt{n} = x$ go continuously to 0.

For a proof closer in spirit to those that follow, note that (26.5) for $n = 2$ gives $|\varphi(t) - (1 - \frac{1}{2}t^2)| \le |t|^3 (|X_k| \le 1)$. Therefore,

$$\left|\varphi\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)\right| \le \frac{|t|^3}{n^{3/2}}. \tag{27.3}$$

Rather than take logarithms, use (27.5) below, which gives (n large)

$$\left|\varphi^n\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)^n\right| \le \frac{|t|^3}{\sqrt{n}} \to 0. \tag{27.4}$$

But of course $(1 - t^2/2n)^n \to e^{-t^2/2}$, which completes the proof for this special case.

Logarithms for complex arguments can be avoided by use of the following simple lemma.

Lemma 1. *Let $z_1, \ldots, z_m$ and $w_1, \ldots, w_m$ be complex numbers of modulus at most* 1; *then*

$$|z_1 \cdots z_m - w_1 \cdots w_m| \le \sum_{k=1}^{m} |z_k - w_k|. \tag{27.5}$$

PROOF. This follows by induction from $z_1 \cdots z_m - w_1 \cdots w_m = (z_1 - w_1)(z_2 \cdots z_m) + w_1(z_2 \cdots z_m - w_2 \cdots w_m)$. ■

Two illustrations of Theorem 27.1:

Example 27.1. In the classical De Moivre-Laplace theorem, X_n takes the values 1 and 0 with probabilities p and $q = 1 - p$, so that $c = p$, and $\sigma^2 = pq$. Here S_n is the number of successes in n Bernoulli trials, and $(S_n - np)/\sqrt{npq} \Rightarrow N$. ■

Example 27.2. Suppose that one wants to estimate the parameter α of an exponential distribution (20.10) on the basis of an independent sample $X_1, \ldots, X_n$. As $n \to \infty$ the sample mean $\bar{X}_n = n^{-1}\sum_{k=1}^{n} X_k$ converges in probability to the mean $1/\alpha$ of the distribution, and hence it is natural to use $1/\bar{X}_n$ to estimate α itself. How good is the estimate? The variance of the exponential distribution being $1/\alpha^2$ (Example 21.3), $\alpha\sqrt{n}\,(\bar{X}_n - 1/\alpha) \Rightarrow N$ by the Lindeberg-Lévy theorem. Thus $\bar{X}_n$ is approximately normally distributed with mean $1/\alpha$ and standard deviation $1/\alpha\sqrt{n}$.

By Skorohod's Theorem 25.6 there exist on a single probability space random variables $\bar{Y}_n$ and Y having the respective distributions of $\bar{X}_n$ and N and satisfying $\alpha\sqrt{n}\,(\bar{Y}_n(\omega) - 1/\alpha) \to Y(\omega)$ for each ω. Now $\bar{Y}_n(\omega) \to 1/\alpha$ and $\alpha^{-1}\sqrt{n}\,(\bar{Y}_n(\omega)^{-1} - \alpha) = \alpha\sqrt{n}\,(\alpha^{-1} - \bar{Y}_n(\omega))/\alpha\bar{Y}_n(\omega) \to -Y(\omega)$. Since $-Y$ has

the distribution of N and $\bar{Y}_n$ has the distribution of $\bar{X}_n$, it follows that

$$\frac{\sqrt{n}}{\alpha}\left(\frac{1}{\bar{X}_n} - \alpha\right) \Rightarrow N;$$

thus $1/\bar{X}_n$ is approximately normally distributed with mean α and standard deviation $\alpha/\sqrt{n}$. In effect, $1/\bar{X}_n$ has been studied through the local linear approximation to the function $1/x$. This is called the *delta method*. ■

The Lindeberg and Lyapounov Theorems

Suppose that for each n

$$X_{n1}, \ldots, X_{nr_n} \tag{27.6}$$

are independent; the probability space for the sequence may change with n. Such a collection is called a *triangular array* of random variables. Put $S_n = X_{n1} + \cdots + X_{nr_n}$. Theorem 27.1 covers the special case in which $r_n = n$ and $X_{nk} = X_k$. Example 6.3 on the number of cycles in a random permutation shows that the idea of triangular array is natural and useful. The central limit theorem for triangular arrays will be applied in Example 27.3 to the same array.

To establish the asymptotic normality of S_n by means of the ideas in the preceding proof requires expanding the characteristic function of each X_{nk} to second-order terms and estimating the remainder. Suppose that the means are 0 and the variances are finite; write

$$E[X_{nk}] = 0, \qquad \sigma_{nk}^2 = E[X_{nk}^2], \qquad s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2. \tag{27.7}$$

The assumption that X_{nk} has mean 0 entails no loss of generality. Assume $s_n^2 > 0$ for large n. A successful remainder estimate is possible under the assumption of the *Lindeberg condition*:

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{|X_{nk}| \geq \epsilon s_n} X_{nk}^2 \, dP = 0 \tag{27.8}$$

for $\epsilon > 0$.

Theorem 27.2. *Suppose that for each n the sequence $X_{n1}, \ldots, X_{nr_n}$ is independent and satisfies* (27.7). *If* (27.8) *holds for all positive ϵ, then $S_n/s_n \Rightarrow N$.*

This theorem contains the preceding one: Suppose that $X_{nk} = X_k$ and $r_n = n$, where the entire sequence $\{X_k\}$ is independent and the X_k all have the same distribution with mean 0 and variance σ^2. Then (27.8) reduces to

$$\lim_{n\to\infty} \frac{1}{\sigma^2} \int_{|X_1| \geq \epsilon\sigma\sqrt{n}} X_1^2 \, dP = 0, \tag{27.9}$$

which holds because $[|X_1| \geq \epsilon\sigma\sqrt{n}\,] \downarrow \varnothing$ as $n \uparrow \infty$.

PROOF OF THE THEOREM. Replacing X_{nk} by X_{nk}/s_n shows that there is no loss of generality in assuming

$$s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 = 1. \tag{27.10}$$

By (26.4_2),

$$\left| e^{itx} - \left(1 + itx - \tfrac{1}{2}t^2x^2\right) \right| \leq \min\left\{ |tx|^2, |tx|^3 \right\}.$$

Therefore, the characteristic function φ_{nk} of X_{nk} satisfies

$$\left| \varphi_{nk}(t) - \left(1 - \tfrac{1}{2}t^2\sigma_{nk}^2\right) \right| \leq E\left[\min\left\{ |tX_{nk}|^2, |tX_{nk}|^3 \right\} \right]. \tag{27.11}$$

Note that the expected value is finite.

For positive ϵ the right side of (27.11) is at most

$$\int_{|X_{nk}|<\epsilon} |tX_{nk}|^3 \, dP + \int_{|X_{nk}|\geq\epsilon} |tX_{nk}|^2 \, dP \leq \epsilon|t|^3\sigma_{nk}^2 + t^2 \int_{|X_{nk}|\geq\epsilon} X_{nk}^2 \, dP.$$

Since the σ_{nk}^2 add to 1 and ϵ is arbitrary, it follows by the Lindeberg condition that

$$\sum_{k=1}^{r_n} \left| \varphi_{nk}(t) - \left(1 - \tfrac{1}{2}t^2\sigma_{nk}^2\right) \right| \to 0 \tag{27.12}$$

for each fixed t. The objective now is to show that

$$\begin{aligned} \prod_{k=1}^{r_n} \varphi_{nk}(t) &= \prod_{k=1}^{r_n} \left(1 - \tfrac{1}{2}t^2\sigma_{nk}^2\right) + o(1) \\ &= \prod_{k=1}^{r_n} e^{-t^2\sigma_{nk}^2/2} + o(1) = e^{-t^2/2} + o(1). \end{aligned} \tag{27.13}$$

For ϵ positive,

$$\sigma_{nk}^2 \le \epsilon^2 + \int_{|X_{nk}|\ge\epsilon} X_{nk}^2\, dP,$$

and so it follows by the Lindeberg condition (recall that s_n is now 1) that

$$\max_{1\le k\le r_n} \sigma_{nk}^2 \to 0. \tag{27.14}$$

For large enough n, $1-\frac{1}{2}t^2\sigma_{nk}^2$ are all between 0 and 1, and by (27.5), $\prod_{k=1}^{r_n}\varphi_{nk}(t)$ and $\prod_{k=1}^{r_n}(1-\frac{1}{2}t^2\sigma_{nk}^2)$ differ by at most the sum in (27.12). This establishes the first of the asymptotic relations in (27.13).

Now (27.5) also implies that

$$\left|\prod_{k=1}^{r_n} e^{-t^2\sigma_{nk}^2/2} - \prod_{k=1}^{r_n}\left(1-\tfrac{1}{2}t^2\sigma_{nk}^2\right)\right| \le \sum_{k=1}^{r_n}\left|e^{-t^2\sigma_{nk}^2/2} - 1 + \tfrac{1}{2}t^2\sigma_{nk}^2\right|.$$

For complex z,

$$|e^z - 1 - z| \le |z|^2 \sum_{k=2}^{\infty}\frac{|z|^{k-2}}{k!} \le |z|^2 e^{|z|}. \tag{27.15}$$

Using this in the right member of the preceding inequality bounds it by $t^4 e^{t^2}\sum_{k=1}^{r_n}\sigma_{nk}^4$; by (27.14) and (27.10), this sum goes to 0, from which the second equality in (27.13) follows. ■

It is shown in the next section (Example 28.4) that if the independent array $\{X_{nk}\}$ satisfies (27.7), and if $S_n/s_n \Rightarrow N$, then the Lindeberg condition holds, *provided* $\max_{k\le r_n}\sigma_{nk}^2/s_n^2 \to 0$. But this converse fails without the extra condition: Take $X_{nk} = X_k$ normal with mean 0 and variance $\sigma_{nk}^2 = \sigma_k^2$, where $\sigma_1^2 = 1$ and $\sigma_n^2 = ns_{n-1}^2$.

Example 27.3. *Goncharov's theorem.* Consider the sum $S_n = \sum_{k=1}^n X_{nk}$ in Example 6.3. Here S_n is the number of cycles in a random permutation on n letters, the X_{nk} are independent, and

$$P[X_{nk} = 1] = \frac{1}{n-k+1} = 1 - P[X_{nk} = 0].$$

The mean m_n is $L_n = \sum_{k=1}^n k^{-1}$, and the variance s_n^2 is $L_n + O(1)$. Lindeberg's condition for $X_{nk} - (n-k+1)^{-1}$ is easily verified because these random variables are bounded by 1.

The theorem gives $(S_n - L_n)/s_n \Rightarrow N$. Now, in fact, $L_n = \log n + O(1)$, and so (see Example 25.8) the sum can be renormalized: $(S_n - \log n)/\sqrt{\log n} \Rightarrow N$. ■

Suppose that the $|X_{nk}|^{2+\delta}$ are integrable for some positive δ and that *Lyapounov's condition*

$$\lim_n \sum_{k=1}^{r_n} \frac{1}{s_n^{2+\delta}} E\left[|X_{nk}|^{2+\delta}\right] = 0 \tag{27.16}$$

holds. Then Lindeberg's condition follows because the sum in (27.8) is bounded by

$$\sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{|X_{nk}| \geq \epsilon s_n} \frac{|X_{nk}|^{2+\delta}}{\epsilon^\delta s_n^\delta}\, dP \leq \frac{1}{\epsilon^\delta} \sum_{k=1}^{r_n} \frac{1}{s_n^{2+\delta}} E\left[|X_{nk}|^{2+\delta}\right].$$

Hence Theorem 27.2 has this corollary:

Theorem 27.3. *Suppose that for each n the sequence $X_{n1}, \ldots, X_{nr_n}$ is independent and satisfies* (27.7). *If* (27.16) *holds for some positive δ, then $S_n/s_n \Rightarrow N$.*

Example 27.4. Suppose that $X_1, X_2, \ldots$ are independent and uniformly bounded and have mean 0. If the variance s_n^2 of $S_n = X_1 + \cdots + X_n$ goes to ∞, then $S_n/s_n \Rightarrow N$: If K bounds the X_n, then

$$\sum_{k=1}^{n} \frac{1}{s_n^3} E\left[|X_k|^3\right] \leq \sum_{k=1}^{n} \frac{KE\left[X_k^2\right]}{s_n^3} = \frac{K}{s_n} \to 0,$$

which is Lyapounov's condition for $\delta = 1$. ■

Example 27.5. Elements are drawn from a population of size n, randomly and with replacement, until the number of distinct elements that have been sampled is r_n, where $1 \leq r_n \leq n$. Let S_n be the drawing on which this first happens. A coupon collector requires S_n purchases to fill out a given portion of the complete set. Suppose that r_n varies with n in such a way that $r_n/n \to \rho$, $0 < \rho < 1$. What is the approximate distribution of S_n?

Let Y_p be the trial on which success first occurs in a Bernoulli sequence with probability p for success: $P[Y_p = k] = q^{k-1}p$, where $q = 1 - p$. Since the moment generating function is $pe^s/(1 - qe^s)$, $E[Y_p] = p^{-1}$ and $\mathrm{Var}[Y_p] = qp^{-2}$. If $k - 1$ distinct items have thus far entered the sample, the waiting time until the next distinct one enters is distributed as Y_p as $p = (n - k + 1)/n$. Therefore, S_n can be represented as $\sum_{k=1}^{r_n} X_{nk}$ for independent summands X_{nk} distributed as $Y_{(n-k+1)/n}$. Since $r_n \sim \rho n$, the mean and variance above give

$$m_n = E[S_n] = \sum_{k=1}^{r_n} \left(1 - \frac{k-1}{n}\right)^{-1} \sim n \int_0^\rho \frac{dx}{1-x}$$

and

$$s_n^2 = \sum_{k=1}^{r_n} \frac{k-1}{n}\left(1 - \frac{k-1}{n}\right)^{-2} \sim n\int_0^\rho \frac{x\,dx}{(1-x)^2}.$$

Lyapounov's theorem applies for $\delta = 2$, and to check (27.16) requires the inequality

$$E\left[\left(Y_p - p^{-1}\right)^4\right] \le Kp^{-4} \tag{27.17}$$

for some K independent of p. A calculation with the moment generating function shows that the left side is in fact $qp^{-4}(1 + 7q + q^2)$. It now follows that

$$\sum_{k=1}^{r_n} E\left[\left(X_{nk} - \frac{n}{n-k+1}\right)^4\right] \le K \sum_{k=1}^{r_n} \left(1 - \frac{k-1}{n}\right)^{-4} \tag{27.18}$$

$$\sim Kn\int_0^\rho \frac{dx}{(1-x)^4}.$$

Since (27.16) follows from this, Theorem 27.3 applies: $(S_n - m_n)/s_n \Rightarrow N$. ■

Dependent Variables*

The assumption of independence in the preceding theorems can be relaxed in various ways. Here a central limit theorem will be proved for sequences in which random variables far apart from one another are nearly independent in a sense to be defined.

For a sequence $X_1, X_2, \ldots$ of random variables, let α_n be a number such that

$$|P(A \cap B) - P(A)P(B)| \le \alpha_n \tag{27.19}$$

for $A \in \sigma(X_1, \ldots, X_k)$, $B \in \sigma(X_{k+n}, X_{k+n+1}, \ldots)$, and $k \ge 1$, $n \ge 1$. Suppose that $\alpha_n \to 0$, the idea being that X_k and X_{k+n} are then approximately independent for large n. In this case the sequence $\{X_n\}$ is said to be *α-mixing*. If the distribution of the random vector $(X_n, X_{n+1}, \ldots, X_{n+j})$ does not depend on n, the sequence is said to be *stationary*.

Example 27.6. Let $\{Y_n\}$ be a Markov chain with finite state space and positive transition probabilities p_{ij}, and suppose that $X_n = f(Y_n)$, where f is some real function on the state space. If the initial probabilities p_i are the stationary ones (see Theorem 8.9), then clearly $\{X_n\}$ is stationary. Moreover, by (8.42), $|p_{ij}^{(n)} - p_j| \le \rho^n$, where $\rho < 1$. By (8.11), $P[Y_1 = i_1, \ldots, Y_k = i_k, Y_{k+n} = j_0, \ldots, Y_{k+n+l} = j_l] = p_{i_1}p_{i_1 i_2} \cdots p_{i_{k-1}i_k}p_{i_k j_0}^{(n)}p_{j_0 j_1} \cdots p_{j_{l-1}j_l}$, which differs

*This topic may be omitted.

from $P[Y_1 = i_1, \ldots, Y_k = i_k]P[Y_{k+n} = j_0, \ldots, Y_{k+n+l} = j_l]$ by at most $p_{i_1}p_{i_1 i_2}\cdots p_{i_{k-1}i_k}\rho^n p_{j_0 j_1}\cdots p_{j_{l-1}j_l}$. It follows by addition that, if s is the number of states, then for sets of the form $A = [(Y_1, \ldots, Y_k) \in H]$ and $B = [(Y_{k+n}, \ldots, Y_{k+n+l}) \in H']$, (27.19) holds with $\alpha_n = s\rho^n$. These sets (for k and n fixed) form fields generating σ-fields which contain $\sigma(X_1, \ldots, X_k)$ and $\sigma(X_{k+n}, X_{k+n+1}, \ldots)$, respectively. For fixed A the set of B satisfying (27.19) is a monotone class, and similarly if A and B are interchanged. It follows by the monotone class theorem (Theorem 3.4) that $\{X_n\}$ is α-mixing with $\alpha_n = s\rho^n$. ■

The sequence is *m-dependent* if $(X_1, \ldots, X_k)$ and $(X_{k+n}, \ldots, X_{k+n+l})$ are independent whenever $n > m$. In this case the sequence is α-mixing with $\alpha_n = 0$ for $n > m$. In this terminology an independent sequence is 0-dependent.

Example 27.7. Let $Y_1, Y_2, \ldots$ be independent and identically distributed, and put $X_n = f(Y_n, \ldots, Y_{n+m})$ for a real function f on R^{m+1}. Then $\{X_n\}$ is stationary and m-dependent. ■

Theorem 27.4. *Suppose that $X_1, X_2, \ldots$ is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $E[X_n] = 0$ and $E[X_n^{12}] < \infty$. If $S_n = X_1 + \cdots + X_n$, then*

$$(27.20) \qquad n^{-1}\operatorname{Var}[S_n] \to \sigma^2 = E[X_1^2] + 2\sum_{k=1}^{\infty} E[X_1 X_{1+k}],$$

where the series converges absolutely. If $\sigma > 0$, then $S_n/\sigma\sqrt{n} \Rightarrow N$.

The conditions $\alpha_n = O(n^{-5})$ and $E[X_n^{12}] < \infty$ are stronger than necessary; they are imposed to avoid technical complications in the proof. The idea of the proof, which goes back to Markov, is this: Split the sum $X_1 + \cdots + X_n$ into alternate blocks of length b_n (the big blocks) and l_n (the little blocks). Namely, let

$$(27.21) \qquad U_{ni} = X_{(i-1)(b_n+l_n)+1} + \cdots + X_{(i-1)(b_n+l_n)+b_n}, \qquad 1 \le i \le r_n,$$

where r_n is the largest integer i for which $(i-1)(b_n + l_n) + b_n < n$. Further, let

$$(27.22) \qquad \begin{aligned} V_{ni} &= X_{(i-1)(b_n+l_n)+b_n+1} + \cdots + X_{i(b_n+l_n)}, \qquad 1 \le i < r_n, \\ V_{nr_n} &= X_{(r_n-1)(b_n+l_n)+b_n+1} + \cdots + X_n. \end{aligned}$$

Then $S_n = \sum_{i=1}^{r_n} U_{ni} + \sum_{i=1}^{r_n} V_{ni}$, and the technique will be to choose the l_n small enough that $\sum_i V_{ni}$ is small in comparison with $\sum_i U_{ni}$ but large enough

that the U_{ni} are nearly independent, so that Lyapounov's theorem can be adapted to prove $\sum_i U_{ni}$ asymptotically normal.

Lemma 2. *If Y is measurable $\sigma(X_1, \ldots, X_k)$ and bounded by C, and if Z is measurable $\sigma(X_{k+n}, X_{k+n+1}, \ldots)$ and bounded by D, then*

$$|E[YZ] - E[Y]E[Z]| \le 4CD\alpha_n. \tag{27.23}$$

PROOF. It is no restriction to take $C = D = 1$ and (by the usual approximation method) to take $Y = \sum_i y_i I_{A_i}$ and $Z = \sum_j z_j I_{B_j}$ simple ($|y_i|, |z_j| \le 1$). If $d_{ij} = P(A_i \cap B_j) - P(A_i)P(B_j)$, the left side of (27.23) is $|\sum_{ij} y_i z_j d_{ij}|$. Take ξ_i to be $+1$ or -1 as $\sum_j z_j d_{ij}$ is positive or not; now take η_j to be $+1$ or -1 as $\sum_i \xi_i d_{ij}$ is positive or not. Then

$$\begin{aligned}\left|\sum_{i,j} y_i z_j d_{ij}\right| &\le \sum_i \left|\sum_j z_j d_{ij}\right| = \sum_i \xi_i \sum_j z_j d_{ij} \\ &\le \sum_j \left|\sum_i \xi_i d_{ij}\right| = \sum_j \eta_j \sum_i \xi_i d_{ij} = \sum_{i,j} \xi_i \eta_j d_{ij}.\end{aligned}$$

Let $A^{(0)}$ [$B^{(0)}$] be the union of the A_i [B_j] for which $\xi_i = +1$ [$\eta_j = +1$], and let $A^{(1)} = \Omega - A^{(0)}$ [$B^{(1)} = \Omega - B^{(0)}$]. Then

$$\sum_{i,j} \xi_i \eta_j d_{ij} \le \sum_{u,v} \left| P(A^{(u)} \cap B^{(v)}) - P(A^{(u)})P(B^{(v)}) \right| \le 4\alpha_n. \qquad \blacksquare$$

Lemma 3. *If Y is measurable $\sigma(X_1, \ldots, X_k)$ and $E[Y^4] \le C$, and if Z is measurable $\sigma(X_{k+n}, X_{k+n+1}, \ldots)$ and $E[Z^4] \le D$, then*

$$|E[YZ] - E[Y]E[Z]| \le 8(1 + C + D)\alpha_n^{1/2}. \tag{27.24}$$

PROOF. Let $Y_0 = YI_{[|Y| \le a]}$, $Y_1 = YI_{[|Y| > a]}$, $Z_0 = ZI_{[|Z| \le a]}$, $Z_1 = ZI_{[|Z| > a]}$. By Lemma 2, $|E[Y_0 Z_0] - E[Y_0]E[Z_0]| \le 4a^2\alpha_n$. Further,

$$\begin{aligned}|E[Y_0 Z_1] - E[Y_0]E[Z_1]| &\le E\left[|Y_0 - E[Y_0]| \cdot |Z_1 - E[Z_1]|\right] \\ &\le 2a \cdot 2E[|Z_1|] \le 4aE\left[|Z_1| \cdot |Z_1/a|^3\right] \le 4D/a^2.\end{aligned}$$

Similary, $|E[Y_1 Z_0] - E[Y_1]E[Z_0]| \le 4C/a^2$. Finally,

$$\begin{aligned}|E[Y_1 Z_1] - E[Y_1]E[Z_1]| &\le \mathrm{Var}^{1/2}[Y_1]\,\mathrm{Var}^{1/2}[Z_1] \le E^{1/2}\left[Y_1^2\right]E^{1/2}\left[Z_1^2\right] \\ &\le E^{1/2}\left[Y_1^4/a^2\right]E^{1/2}\left[Z_1^4/a^2\right] \le C^{1/2}D^{1/2}/a^2.\end{aligned}$$

Adding these inequalities gives $4a^2\alpha_n + 4(C + D)a^{-2} + C^{1/2}D^{1/2}a^{-2}$ as a

bound for the left side of (27.24). Take $a = \alpha_n^{-1/4}$ and observe that $4 + 4(C + D) + C^{1/2}D^{1/2} \le 4 + 4(C^{1/2} + D^{1/2})^2 \le 4 + 8(C + D)$. ■

PROOF OF THEOREM 27.4. By Lemma 3, $|E[X_1X_{1+n}]| \le 8(1 + 2E[X_1^4])\alpha_n^{1/2} = O(n^{-5/2})$, and so the series in (27.20) converges absolutely. If $\rho_k = E[X_1X_{1+k}]$, then by stationary $E[S_n^2] = n\rho_0 + 2\sum_{k=1}^{n-1}(n-k)\rho_k$ and therefore $|\sigma^2 - n^{-1}E[S_n^2]| \le 2\sum_{k=n}^{\infty}|\rho_k| + 2n^{-1}\sum_{i=1}^{n-1}\sum_{k=i}^{\infty}|\rho_k|$; hence (27.20).

By stationarity again,

$$E[S_n^4] \le 4!n\sum\left|E[X_1X_{1+i}X_{1+i+j}X_{1+i+j+k}]\right|,$$

where the indices in the sum are constrained by $i, j, k \ge 0$ and $i + j + k < n$. By Lemma 3 the summand is at most

$$8\left(1 + E[X_1^4] + E[X_{1+i}^4X_{1+i+j}^4X_{1+i+j+k}^4]\right)\alpha_i^{1/2},$$

which is at most[†]

$$8\left(1 + E[X_1^4] + E[X_1^{12}]\right)\alpha_i^{1/2} = K_1\alpha_i^{1/2}.$$

Similarly, $K_1\alpha_k^{1/2}$ is a bound. Hence

$$E[S_n^4] \le 4!n^2 \sum_{\substack{i,k\ge 0\\ i+k<n}} K_1 \min\{\alpha_i^{1/2}, \alpha_k^{1/2}\}$$

$$\le K_2n^2 \sum_{0\le i\le k} \alpha_k^{1/2} = K_2n^2 \sum_{k=0}^{\infty}(k+1)\alpha_k^{1/2}.$$

Since $\alpha_k = O(k^{-5})$, the series here converges, and therefore

$$E[S_n^4] \le Kn^2 \tag{27.25}$$

for some K independent of n.

Let $b_n = \lfloor n^{3/4} \rfloor$ and $l_n = \lfloor n^{1/4} \rfloor$. If r_n is the largest integer i such that $(i-1)(b_n + l_n) + b_n < n$, then

$$b_n \sim n^{3/4}, \qquad l_n \sim n^{1/4}, \qquad r_n \sim n^{1/4}. \tag{27.26}$$

Consider the random variables (27.21) and (27.22). By (27.25), (27.26), and

[†] $E|XYZ| \le E^{1/3}|X|^3 \cdot E^{2/3}|YZ|^{3/2} \le E^{1/3}|X|^3 \cdot E^{1/3}|Y|^3 \cdot E^{1/3}|Z|^3$.

stationarity,

$$P\left[\left|\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{r_n-1}V_{ni}\right|\geq\epsilon\right]\leq\sum_{i=1}^{r_n-1}P\left[|V_{ni}|\geq\frac{\epsilon\sigma\sqrt{n}}{r_n}\right]$$
$$\leq\frac{r_n^4}{\epsilon^4\sigma^4n^2}r_nKl_n^2\sim\frac{K}{\epsilon^4\sigma^4n^{1/4}}\to 0;$$

(27.25) and (27.26) also give

$$P\left[\frac{1}{\sigma\sqrt{n}}|V_{nr_n}|\geq\epsilon\right]\leq\frac{K(b_n+l_n)^2}{\epsilon^4\sigma^4n^2}\sim\frac{K}{\epsilon^4\sigma^4n^{1/2}}\to 0.$$

Therefore, $\sum_{i=1}^{r_n}V_{ni}/\sigma\sqrt{n}\Rightarrow 0$, and by Theorem 25.4 it suffices to prove that $\sum_{i=1}^{r_n}U_{ni}/\sigma\sqrt{n}\Rightarrow N$.

Let U'_{ni}, $1\leq i\leq r_n$, be independent random variables having the distribution common to the U_{ni}. By Lemma 2 extended inductively the characteristic functions of $\sum_{i=1}^{r_n}U_{ni}/\sigma\sqrt{n}$ and of $\sum_{i=1}^{r_n}U'_{ni}/\sigma\sqrt{n}$ differ by at most† $16r_n\alpha_{l_n}$. Since $\alpha_n=O(n^{-5})$, this difference is $O(n^{-1})$ by (27.26).

The characteristic function of $\sum_{i=1}^{r_n}U_{ni}/\sigma\sqrt{n}$ will thus approach $e^{-t^2/2}$ if that of $\sum_{i=1}^{r_n}U'_{ni}/\sigma\sqrt{n}$ does. It therefore remains only to show that $\sum_{i=1}^{r_n}U'_{ni}/\sigma\sqrt{n}\Rightarrow N$. Now $E[|U'_{ni}|^2]=E[U_{n1}^2]\sim b_n\sigma^2$ by (27.20). Further, $E[|U'_{ni}|^4]\leq Kb_n^2$ by (27.25). Lyapounov's condition (27.16) for $\delta=2$ therefore follows because

$$\frac{r_nE\left[|U'_{n1}|^4\right]}{\left(r_nE\left[|U'_{n1}|^2\right]\right)^2}\sim\frac{E\left[|U'_{n1}|^4\right]}{r_nb_n^2\sigma^4}\leq\frac{K}{r_n\sigma^4}\to 0.$$ ■

Example 27.8. Let $\{Y_n\}$ be the stationary Markov process of Example 27.6. Let f be a function on the state space, put $m=\sum_i p_if(i)$, and define $X_n=f(Y_n)-m$. Then $\{X_n\}$ satisfies the conditions of Theorem 27.4. If $\beta_{ij}=\delta_{ij}p_i-p_ip_j+2p_i\sum_{k=1}^{\infty}(p_{ij}^{(k)}-p_j)$, then the σ^2 in (27.20) is $\sum_{ij}\beta_{ij}(f(i)-m)(f(j)-m)$, and $\sum_{k=1}^{n}f(Y_k)$ is approximately normally distributed with mean nm and standard deviation $\sigma\sqrt{n}$.

If $f(i)=\delta_{i_0i}$, then $\sum_{k=1}^{n}f(Y_k)$ is the number of passages through the state i_0 in the first n steps of the process. In this case $m=p_{i_0}$ and $\sigma^2=p_{i_0}(1-p_{i_0})+2p_{i_0}\sum_{k=1}^{\infty}(p_{i_0i_0}^{(k)}-p_{i_0})$. ■

Example 27.9. If the X_n are stationary and m-dependent and have mean 0, Theorem 27.4 applies and $\sigma^2=E[X_1^2]+2\sum_{k=1}^{m}E[X_1X_{1+k}]$. Example 27.7 is a case in point. Taking $m=1$ and $f(x,y)=x-y$ in that example gives an instance where $\sigma^2=0$. ■

†The 4 in (27.23) has become 16 to allow for splitting into real and imaginary parts.

PROBLEMS

27.1. Prove Theorem 23.2 by means of characteristic functions. *Hint*: Use (27.5) to compare the characteristic function of $\sum_{k=1}^{r_n} Z_{nk}$ with $\exp[\sum_k p_{nk}(e^{it} - 1)]$.

27.2. If $\{X_n\}$ is independent and the X_n all have the same distribution with finite first moment, then $n^{-1}S_n \to E[X_1]$ with probability 1 (Theorem 22.1), so that $n^{-1}S_n \Rightarrow E[X_1]$. Prove the latter fact by characteristic functions. *Hint*: Use (27.5).

27.3. For a Poisson variable Y_λ with mean λ, show that $(Y_\lambda - \lambda)/\sqrt{\lambda} \Rightarrow N$ as $\lambda \to \infty$. Show that (22.3) fails for $t = 1$.

27.4. Suppose that $|X_{nk}| \leq M_n$ with probability 1 and $M_n/s_n \to 0$. Verify Lyapounov's condition and then Lindeberg's condition.

27.5. Suppose that the random variables in any single row of the triangular array are identically distributed. To what do Lindeberg's and Lyapounov's conditions reduce?

27.6. Suppose that $Z_1, Z_2, \ldots$ are independent and identically distributed with mean 0 and variance 1, and suppose that $X_{nk} = \sigma_{nk} Z_k$. Write down the Lindeberg condition and show that it holds if $\max_{k \leq r_n} \sigma_{nk}^2 = o(\sum_{k=1}^{r_n} \sigma_{nk}^2)$.

27.7. Construct an example where Lindeberg's condition holds but Lyapounov's does not.

27.8. 22.9↑ Prove a central limit theorem for the number R_n of records up to time n.

27.9. 6.3↑ Let S_n be the number of inversions in a random permutation on n letters. Prove a central limit theorem for S_n.

27.10. *The δ-method.* Suppose that Theorem 27.1 applies to $\{X_n\}$, so that $\sqrt{n}\,\sigma^{-1}(\bar{X}_n - c) \Rightarrow N$, where $\bar{X}_n = n^{-1}\sum_{k=1}^n X_k$. Use Theorem 25.6 as in Example 27.2 to show that, if $f(x)$ has a nonzero derivative at c, then $\sqrt{n}\,(f(\bar{X}_n) - f(c))/\sigma|f'(c)| \Rightarrow N$: $\bar{X}_n$ is approximately normal with mean c and standard deviation $\sigma/\sqrt{n}$, and $f(\bar{X}_n)$ is approximately normal with mean $f(c)$ and standard deviation $|f'(c)|\sigma/\sqrt{n}$. Example 27.2 is the case $f(x) = 1/x$.

27.11. Suppose independent X_n have density $|x|^{-3}$ outside $(-1, +1)$. Show that $(n \log n)^{-1/2} S_n \Rightarrow N$.

27.12. There can be asymptotic normality even if there are no moments at all. Construct a simple example.

27.13. Let $d_n(\omega)$ be the dyadic digits of a point ω drawn at random from the unit interval. For a k-tuple $(u_1, \ldots, u_k)$ of 0's and 1's, let $N_n(u_1, \ldots, u_k; \omega)$ be the number of $m \leq n$ for which $(d_m(\omega), \ldots, d_{m+k-1}(\omega)) = (u_1, \ldots, u_k)$. Prove a central limit theorem for $N_n(u_1, \ldots, u_k; \omega)$. (See Problem 6.12.)

27.14. *The central limit theorem for a random number of summands.* Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with mean 0 and variance σ^2, and let $S_n = X_1 + \cdots + X_n$. For each positive t, let ν_t be a random variable assuming positive integers as values; it need not be independent of the X_n. Suppose that there exist positive constants a_t and θ such that

$$a_t \to \infty, \qquad \frac{\nu_t}{a_t} \Rightarrow \theta$$

as $t \to \infty$. Show by the following steps that

$$\text{(27.27)} \qquad \frac{S_{\nu_t}}{\sigma\sqrt{\nu_t}} \Rightarrow N, \qquad \frac{S_{\nu_t}}{\sigma\sqrt{\theta a_t}} \Rightarrow N.$$

(a) Show that it may be assumed that $\theta = 1$ and the a_t are integers.

(b) Show that it suffices to prove the second relation in (27.27).

(c) Show that it suffices to prove $(S_{\nu_t} - S_{a_t})/\sqrt{a_t} \Rightarrow 0$.

(d) Show that

$$P\left[|S_{\nu_t} - S_{a_t}| \geq \epsilon\sqrt{a_t}\right] \leq P\left[|\nu_t - a_t| \geq \epsilon^3 a_t\right] + P\left[\max_{|k - a_t| \leq \epsilon^3 a_t} |S_k - S_{a_t}| \geq \epsilon\sqrt{a_t}\right],$$

and conclude from Kolmogorov's inequality that the last probability is at most $2\epsilon\sigma^2$.

27.15. 21.21 23.10 23.14↑ *A central limit theorem in renewal theory.* Let $X_1, X_2, \ldots$ be independent, identically distributed positive random variables with mean m and variance σ^2, and as in Problem 23.10 let N_t be the maximum n for which $S_n \leq t$. Prove by the following steps that

$$\frac{N_t - tm^{-1}}{\sigma t^{1/2} m^{-3/2}} \Rightarrow N.$$

(a) Show by the results in Problems 21.21 and 23.10 that $(S_{N_t} - t)/\sqrt{t} \Rightarrow 0$.

(b) Show that it suffices to prove that

$$\frac{N_t - S_{N_t} m^{-1}}{\sigma t^{1/2} m^{-3/2}} = \frac{-(S_{N_t} - mN_t)}{\sigma t^{1/2} m^{-1/2}} \Rightarrow N.$$

(c) Show (Problem 23.10) that $N_t/t \Rightarrow m^{-1}$, and apply the theorem in Problem 27.14.

27.16. Show by partial integration that

$$\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-u^2/2}\,du \sim \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2} \tag{27.28}$$

as $x \to \infty$.

27.17. ↑ Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with mean 0 and variance 1, and suppose that $a_n \to \infty$. Formally combine the central limit theorem and (27.28) to obtain

$$P\left[S_n \geq a_n\sqrt{n}\right] \sim \frac{1}{\sqrt{2\pi}}\frac{1}{a_n}e^{-a_n^2/2} = e^{-a_n^2(1+\zeta_n)/2}, \tag{27.29}$$

where $\zeta_n \to 0$ if $a_n \to \infty$. For a case in which this does hold, see Theorem 9.4.

27.18. 21.2↑ *Stirling's formula*. Let $S_n = X_1 + \cdots + X_n$, where the X_n are independent and each has the Poisson distribution with parameter 1. Prove successively:

(a) $E\left[\left(\frac{S_n - n}{\sqrt{n}}\right)^-\right] = e^{-n}\sum_{k=0}^{n}\left(\frac{n-k}{\sqrt{n}}\right)\frac{n^k}{k!} = \frac{n^{n+(1/2)}e^{-n}}{n!}$.

(b) $\left(\frac{S_n - n}{\sqrt{n}}\right)^- \Rightarrow N^-$.

(c) $E\left[\left(\frac{S_n - n}{\sqrt{n}}\right)^-\right] \to E[N^-] = \frac{1}{\sqrt{2\pi}}$.

(d) $n! \sim \sqrt{2\pi}\,n^{n+(1/2)}e^{-n}$.

27.19. Let $l_n(\omega)$ be the length of the run of 0's starting at the nth place in the dyadic expansion of a point ω drawn at random from the unit interval; see Example 4.1.

(a) Show that $l_1, l_2, \ldots$ is an α-mixing sequence, where $\alpha_n = 4/2^n$.

(b) Show that $\sum_{k=1}^{n} l_k$ is approximately normally distributed with mean n and variance $6n$.

27.20. Prove under the hypotheses of Theorem 27.4 that $S_n/n \to 0$ with probability 1. *Hint*: Use (27.25).

27.21. 26.1 26.29↑ Let $X_1, X_2, \ldots$ be independent and identically distributed, and suppose that the distribution common to the X_n is supported by $[0, 2\pi]$ and is not a lattice distribution. Let $S_n = X_1 + \cdots + X_n$, where the sum is reduced modulo 2π. Show that $S_n \Rightarrow U$, where U is uniformly distributed over $[0, 2\pi]$.

SECTION 28. INFINITELY DIVISIBLE DISTRIBUTIONS*

Suppose that Z_λ has the Poisson distribution with parameter λ and that $X_{n1}, \ldots, X_{nn}$ are independent and $P[X_{nk} = 1] = \lambda/n$, $P[X_{nk} = 0] = 1 - \lambda/n$. According to Example 25.2, $X_{n1} + \cdots + X_{nn} \Rightarrow Z_\lambda$. This contrasts with the central limit theorem, in which the limit law is normal. What is the class of all possible limit laws for independent triangular arrays? A suitably restricted form of this question will be answered here.

Vague Convergence

The theory requires two preliminary facts about convergence of measures. Let μ_n and μ be finite measures on $(R^1, \mathscr{R}^1)$. If $\mu_n(a, b] \to \mu(a, b]$ for every finite interval for which $\mu\{a\} = \mu\{b\} = 0$, then μ_n *converges vaguely* to μ, written $\mu_n \to_v \mu$. If μ_n and μ are probability measures, it is not hard to see that this is equivalent to weak convergence: $\mu_n \Rightarrow \mu$. On the other hand, if μ_n is a unit mass at n and $\mu(R^1) = 0$, then $\mu_n \to_v \mu$, but $\mu_n \Rightarrow \mu$ makes no sense, because μ is not a probability measure.

The first fact needed is this: *Suppose that* $\mu_n \to_v \mu$ *and*

$$\sup_n \mu_n(R^1) < \infty; \tag{28.1}$$

then

$$\int f d\mu_n \to \int f d\mu \tag{28.2}$$

for every continuous real f *that vanishes at* $\pm\infty$ *in the sense that* $\lim_{|x| \to \infty} f(x) = 0$. Indeed, choose M so that $\mu(R^1) < M$ and $\mu_n(R^1) < M$ for all n. Given ϵ, choose a and b so that $\mu\{a\} = \mu\{b\} = 0$ and $|f(x)| < \epsilon/M$ if $x \notin A = (a, b]$. Then $|\int_{A^c} f d\mu_n| < \epsilon$ and $|\int_{A^c} f d\mu| < \epsilon$. If $\mu(A) > 0$, define $\nu(B) = \mu(B \cap A)/\mu(A)$ and $\nu_n(B) = \mu_n(B \cap A)/\mu_n(A)$. It is easy to see that $\nu_n \Rightarrow \nu$, so that $\int f d\nu_n \to \int f d\nu$. But then $|\int_A f d\mu_n - \int_A f d\mu| < \epsilon$ for large n, and hence $|\int f d\mu_n - \int f d\mu| < 3\epsilon$ for large n. If $\mu(A) = 0$, then $\int_A f d\mu_n \to 0$, and the argument is even simpler.

The other fact needed below is this: *If* (28.1) *holds, then there is a subsequence* $\{\mu_{n_k}\}$ *and a finite measure* μ *such that* $\mu_{n_k} \to_v \mu$ *as* $k \to \infty$. Indeed, let $F_n(x) = \mu_n(-\infty, x]$. Since the F_n are uniformly bounded because of (28.1), the proof of Helly's theorem shows there exists a subsequence $\{F_{n_k}\}$ and a bounded, nondecreasing, right-continuous function F such that $\lim_k F_{n_k}(x) = F(x)$ at continuity points x of F. If μ is the measure for which $\mu(a, b] = F(b) - F(a)$ (Theorem 12.4), then clearly $\mu_{n_k} \to_v \mu$.

The Possible Limits

Let $X_{n1}, \ldots, X_{nr_n}$, $n = 1, 2, \ldots$, be a triangular array as in the preceding section. The random variables in each row are independent, the means are 0,

*This section may be omitted.

and the variances are finite:

$$E[X_{nk}] = 0, \qquad \sigma_{nk}^2 = E[X_{nk}^2], \qquad s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2. \tag{28.3}$$

Assume $s_n^2 > 0$ and put $S_n = X_{n1} + \cdots + X_{nr_n}$. Here it will be assumed that the total variance is bounded:

$$\sup_n s_n^2 < \infty. \tag{28.4}$$

In order that the X_{nk} be small compared with S_n, assume that

$$\lim_n \max_{k \le r_n} \sigma_{nk}^2 = 0. \tag{28.5}$$

The arrays in the preceding section were normalized by replacing X_{nk} by X_{nk}/s_n. This has the effect of replacing s_n by 1, in which case of course (28.4) holds, and (28.5) is the same thing as $\max_k \sigma_{nk}^2/s_n^2 \to 0$.

A distribution function F is *infinitely divisible* if for each n there is a distribution function F_n such that F is the n-fold convolution $F_n * \cdots * F_n$ (n copies) of F_n. The class of possible limit laws will turn out to consist of the infinitely divisible distributions with mean 0 and finite variance.[†] It will be possible to exhibit the characteristic functions of these laws in an explicit way.

Theorem 28.1. *Suppose that*

$$\varphi(t) = \exp \int_{R^1} \left(e^{itx} - 1 - itx\right) \frac{1}{x^2} \mu(dx), \tag{28.6}$$

where μ is a finite measure. Then φ is the characteristic function of an infinitely divisible distribution with mean 0 and variance $\mu(R^1)$.

By (26.4_2), the integrand in (28.6) converges to $-t^2/2$ as $x \to 0$; take this as its value at $x = 0$. By (26.4_1), the integrand is at most $t^2/2$ in modulus and so is integrable.

The formula (28.6) is the *canonical representation* of φ, and μ is the *canonical measure*.

Before proceeding to the proof, consider three examples.

Example 28.1. If μ consists of a mass of σ^2 at the origin, (28.6) is $e^{-\sigma^2 t^2/2}$, the characteristic function of a centered normal distribution F. It is certainly infinitely divisible—take F_n normal with variance σ^2/n. ■

[†]There do exist infinitely divisible distributions without moments (see Problems 28.3 and 28.4), but they do not figure in the theory of this section.

Example 28.2. Suppose that μ consists of a mass of λx^2 at $x \neq 0$. Then (28.6) is $\exp \lambda(e^{itx} - 1 - itx)$; but this is the characteristic function of $x(Z_\lambda - \lambda)$, where Z_λ has the Poisson distribution with mean λ. Thus (28.6) is the characteristic function of a distribution function F, and F is infinitely divisible—take F_n to be the distribution function of $x(Z_{\lambda/n} - \lambda/n)$. ■

Example 28.3. If $\varphi_j(t)$ is given by (28.6) with μ_j for the measure, and if $\mu = \sum_{j=1}^k \mu_j$, then (28.6) is $\varphi_1(t) \ldots \varphi_k(t)$. It follows by the preceding two examples that (28.6) is a characteristic function if μ consists of finitely many point masses. It is easy to check in the preceding two examples that the distribution corresponding to $\varphi(t)$ has mean 0 and variance $\mu(R^1)$, and since the means and variances add, the same must be true in the present example. ■

PROOF OF THEOREM 28.1. Let μ_k have mass $\mu(j2^{-k}, (j+1)2^{-k}]$ at $j2^{-k}$ for $j = 0, \pm 1, \ldots, \pm 2^{2k}$. Then $\mu_k \to_v \mu$. As observed in Example 28.3, if $\varphi_k(t)$ is (28.6) with μ_k in place of μ, then φ_k is a characteristic function. For each t the integrand in (28.6) vanishes at $\pm\infty$; since $\sup_k \mu_k(R^1) < \infty$, $\varphi_k(t) \to \varphi(t)$ follows (see (28.2)). By Corollary 2 to Theorem 26.3, $\varphi(t)$ is itself a characteristic function. Further, the distribution corresponding to $\varphi_k(t)$ has second moment $\mu_k(R^1)$, and since this is bounded, it follows (Theorem 25.11) that the distribution corresponding to $\varphi(t)$ has a finite second moment. Differentiation (use Theorem 16.8) shows that the mean is $\varphi'(0) = 0$ and the variance is $-\varphi''(0) = \mu(R^1)$. Thus (28.6) is always the characteristic function of a distribution with mean 0 and variance $\mu(R^1)$.

If $\psi_n(t)$ is (28.6) with μ/n in place of μ, then $\varphi(t) = \psi_n^n(t)$, so that the distribution corresponding to $\varphi(t)$ is indeed infinitely divisible. ■

The representation (28.6) shows that the normal and Poisson distributions are special cases in a very large class of infinitely divisible laws.

Theorem 28.2. *Every infinitely divisible distribution with mean* 0 *and finite variance is the limit law of* S_n *for some independent triangular array satisfying* (28.3), (28.4), *and* (28.5). ■

The proof requires this preliminary result:

Lemma. *If* X *and* Y *are independent and* $X + Y$ *has a second moment, then* X *and* Y *have second moments as well.*

PROOF. Since $X^2 + Y^2 \leq (X+Y)^2 + 2|XY|$, it suffices to prove $|XY|$ integrable, and by Fubini's theorem applied to the joint distribution of X and Y it suffices to prove $|X|$ and $|Y|$ individually integrable. Since $|Y| \leq |x| + |x+Y|$, $E[|Y|] = \infty$ would imply $E[|x+Y|] = \infty$ for each x; by Fubini's

theorem again $E[|Y|] = \infty$ would therefore imply $E[|X + Y|] = \infty$, which is impossible. Hence $E[|Y|] < \infty$, and similarly $E[|X|] < \infty$. ■

PROOF OF THEOREM 28.2. Let F be infinitely divisible with mean 0 and variance σ^2. If F is the n-fold convolution of F_n, then by the lemma (extended inductively) F_n has finite mean and variance, and these must be 0 and σ^2/n. Take $r_n = n$ and take $X_{n1}, \ldots, X_{nn}$ independent, each with distribution function F_n. ■

Theorem 28.3. *If F is the limit law of S_n for an independent triangular array satisfying* (28.3), (28.4), *and* (28.5), *then F has characteristic function of the form* (28.6) *for some finite measure μ.*

PROOF. The proof will yield information making it possible to identify the limit. Let $\varphi_{nk}(t)$ be the characteristic function of X_{nk}. The first step is to prove that

$$\prod_{k=1}^{r_n} \varphi_{nk}(t) - \exp \sum_{k=1}^{r_n} (\varphi_{nk}(t) - 1) \to 0 \tag{28.7}$$

for each t. Since $|z| \le 1$ implies that $|e^{z-1}| = e^{\operatorname{Re} z - 1} \le 1$, it follows by (27.5) that the difference $\delta_n(t)$ in (28.7) satisfies $|\delta_n(t)| \le \sum_{k=1}^{r_n} |\varphi_{nk}(t) - \exp(\varphi_{nk}(t) - 1)|$. Fix t. If $\varphi_{nk}(t) - 1 = \theta_{nk}$, then $|\theta_{nk}| \le t^2\sigma_{nk}^2/2$, and it follows by (28.4) and (28.5) that $\max_k |\theta_{nk}| \to 0$ and $\sum_k |\theta_{nk}| = O(1)$. Therefore, for sufficiently large n, $|\delta_n(t)| \le \sum_k |1 + \theta_{nk} - e^{\theta_{nk}}| \le e^2 \sum_k |\theta_{nk}|^2 \le e^2 \max_k |\theta_{nk}| \cdot \sum_k |\theta_{nk}|$ by (27.15). Hence (28.7).

If F_{nk} is the distribution function of X_{nk}, then

$$\begin{aligned} \sum_{k=1}^{r_n} (\varphi_{nk}(t) - 1) &= \sum_{k=1}^{r_n} \int_{R^1} (e^{itx} - 1)\, dF_{nk}(x) \\ &= \sum_{k=1}^{r_n} \int_{R^1} (e^{itx} - 1 - itx)\, dF_{nk}(x). \end{aligned}$$

Let μ_n be the finite measure satisfying

$$\mu_n(-\infty, x] = \sum_{k=1}^{r_n} \int_{y \le x} y^2\, dF_{nk}(y), \tag{28.8}$$

and put

$$\varphi_n(t) = \exp \int_{R^1} (e^{itx} - 1 - itx) \frac{1}{x^2} \mu_n(dx). \tag{28.9}$$

Then (28.7) can be written

$$\prod_{k=1}^{r_n} \varphi_{nk}(t) - \varphi_n(t) \to 0. \tag{28.10}$$

By (28.8), $\mu_n(R^1) = s_n^2$, and this is bounded by assumption. Thus (28.1) holds, and some subsequence $\{\mu_{n_u}\}$ converges vaguely to a finite measure μ. Since the integrand in (28.9) vanishes at $\pm\infty$, $\varphi_{n_u}(t)$ converges to (28.6). But, of course, $\lim_n \varphi_n(t)$ must coincide with the characteristic function of the limit law F, which exists by hypothesis. Thus F must have characteristic function of the form (28.6). ■

Theorems 28.1, 28.2, and 28.3 together show that the possible limit laws are exactly the infinitely divisible distributions with mean 0 and finite variance, and they give explicitly the form the characteristic functions of such laws must have.

Characterizing the Limit

Theorem 28.4. *Suppose that F has characteristic function* (28.6) *and that an independent triangular array satisfies* (28.3), (28.4), *and* (28.5). *Then S_n has limit law F if and only if $\mu_n \to_v \mu$, where μ_n is defined by* (28.8).

PROOF. Since (28.7) holds as before, S_n has limit law F if and only if $\varphi_n(t)$ (defined by (28.9)) converges for each t to $\varphi(t)$ (defined by (28.6)). If $\mu_n \to_v \mu$, then $\varphi_n(t) \to \varphi(t)$ follows because the integrand in (28.9) and (28.6) vanishes at $\pm\infty$ and because (28.1) follows from (28.4).

Now suppose that $\varphi_n(t) \to \varphi(t)$. Since $\mu_n(R^1) = s_n^2$ is bounded, each subsequence $\{\mu_{n_u}\}$ contains a further subsequence $\{\mu_{n_{u(j)}}\}$ converging vaguely to some ν. If it can be shown that ν necessarily coincides with μ, it will follow by the usual argument that $\mu_n \to_v \mu$. But by the definition (28.9) of $\varphi_n(t)$, it follows that $\varphi(t)$ must coincide with $\psi(t) = \exp \int_{R^1}(e^{itx} - 1 - itx)x^{-2}\nu(dx)$. Now $\varphi'(t) = i\varphi(t)\int_{R^1}(e^{itx} - 1)x^{-1}\mu(dx)$, and similarly for $\psi'(t)$. Hence $\varphi(t) = \psi(t)$ implies that $\int_{R^1}(e^{itx} - 1)x^{-1}\nu(dx) = \int_{R^1}(e^{itx} - 1)x^{-1}\mu(dx)$. A further differentiation gives $\int_{R^1} e^{itx}\mu(dx) = \int_{R^1} e^{itx}\nu(dx)$. This implies that $\mu(R^1) = \nu(R^1)$, and so $\mu = \nu$ by the uniqueness theorem for characteristic functions. ■

Example 28.4. *The normal case.* According to the theorem, $S_n \Rightarrow N$ if and only if μ_n converges vaguely to a unit mass at 0. If $s_n^2 = 1$, this holds if and only if $\sum_{k=1}^{r_n} \int_{|x| \geq \epsilon} x^2\, dF_{nk}(x) \to 0$, which is exactly Lindeberg's condition. ■

Example 28.5. *The Poisson case.* Let $Z_{n1}, \ldots, Z_{nr_n}$ be an independent triangular array, and suppose $X_{nk} = Z_{nk} - m_{nk}$ satisfies the conditions of the

theorem, where $m_{nk} = E[Z_{nk}]$. If Z_λ has the Poisson distribution with parameter λ, then $\sum_k X_{nk} \Rightarrow Z_\lambda - \lambda$ if and only if μ_n converges vaguely to a mass of λ at 1 (see Example 28.2). If $s_n^2 \to \lambda$, the requirement is $\mu_n[1-\epsilon, 1+\epsilon] \to \lambda$, or

$$\sum_k \int_{|Z_{nk}-m_{nk}-1|>\epsilon} (Z_{nk}-m_{nk})^2 \, dP \to 0 \tag{28.11}$$

for positive ϵ. If s_n^2 and $\sum_k m_{nk}$ both converge to λ, (28.11) is a necessary and sufficient condition for $\sum_k Z_{nk} \Rightarrow Z_\lambda$. The conditions are easily checked under the hypotheses of Theorem 23.2: Z_{nk} assumes the values 1 and 0 with probabilities p_{nk} and $1-p_{nk}$, $\sum_k p_{nk} \to \lambda$, and $\max_k p_{nk} \to 0$. ■

PROBLEMS

28.1. Show that $\mu_n \to_v \mu$ implies $\mu(R^1) \le \liminf_n \mu_n(R^1)$. Thus in vague convergence mass can "escape to infinity" but mass cannot "enter from infinity."

28.2. (a) Show that $\mu_n \to_v \mu$ if and only if (28.2) holds for every continuous f with bounded support.

(b) Show that if $\mu_n \to_v \mu$ but (28.1) does not hold, then there is a continuous f vanishing at $\pm\infty$ for which (28.2) does not hold.

28.3. 23.7↑ Suppose that $N, Y_1, Y_2, \ldots$ are independent, the Y_n have a common distribution function F, and N has the Poisson distribution with mean α. Then $S = Y_1 + \cdots + Y_N$ has the *compound Poisson distribution*.

(a) Show that the distribution of S is infinitely divisible. Note that S may not have a mean.

(b) The distribution function of S is $\sum_{n=0}^{\infty} e^{-\alpha}\alpha^n F^{n*}(x)/n!$, where F^{n*} is the n-fold convolution of F (a unit jump at 0 for $n = 0$). The characteristic function of S is $\exp \alpha \int_{-\infty}^{\infty} (e^{itx} - 1)\, dF(x)$.

(c) Show that, if F has mean 0 and finite variance, then the canonical measure μ in (28.6) is specified by $\mu(A) = \alpha \int_A x^2 \, dF(x)$.

28.4. (a) Let ν be a finite measure, and define

$$\varphi(t) = \exp\left[i\gamma t + \int_{-\infty}^{\infty} \left(e^{itx} - 1 - \frac{itx}{1+x^2}\right) \frac{1+x^2}{x^2} \nu(dx)\right], \tag{28.12}$$

where the integrand is $-t^2/2$ at the origin. Show that this is the characteristic function of an infinitely divisible distribution.

(b) Show that the Cauchy distribution (see the table on p. 348) is the case where $\gamma = 0$ and ν has density $\pi^{-1}(1+x^2)^{-1}$ with respect to Lebesgue measure.

28.5. Show that the Cauchy, exponential, and gamma (see (20.47)) distributions are infinitely divisible.

28.6. Find the canonical representation (28.6) of the exponential distribution with mean 1:

(a) The characteristic function is $\int_0^\infty e^{itx}e^{-x}\,dx = (1-it)^{-1} = \varphi(t)$.

(b) Show that (use the principal branch of the logarithm or else operate formally for the moment) $d(\log \varphi(t))/dt = i\varphi(t) = i\int_0^\infty e^{itx}e^{-x}\,dx$. Integrate with respect to t to obtain

$$\frac{1}{1-it} = \exp\int_0^\infty \left(e^{itx}-1\right)\frac{e^{-x}}{x}\,dx. \tag{28.13}$$

Verify (28.13) after the fact by showing that the ratio of the two sides has derivative 0.

(c) Multiply (28.13) by e^{-it} to center the exponential distribution at its mean: The canonical measure μ has density xe^{-x} over $(0,\infty)$.

28.7. ↑ If X and Y are independent and each has the exponential density e^{-x}, then $X-Y$ has the double exponential density $\frac{1}{2}e^{-|x|}$ (see the table on p. 348). Show that its characteristic function is

$$\frac{1}{1+t^2} = \exp\int_{-\infty}^\infty \left(e^{itx}-1-itx\right)\frac{1}{x^2}|x|e^{-|x|}\,dx.$$

28.8. ↑ Suppose $X_1, X_2, \ldots$ are independent and each has the double exponential density. Show that $\sum_{n=1}^\infty X_n/n$ converges with probability 1. Show that the distribution of the sum is infinitely divisible and that its canonical measure has density $|x|e^{-|x|}/(1-e^{-|x|}) = \sum_{n=1}^\infty |x|e^{-|nx|}$.

28.9. 26.8↑ Show that for the gamma density $e^{-x}x^{u-1}/\Gamma(u)$ the canonical measure has density uxe^{-x} over $(0,\infty)$.

The remaining problems require the notion of a *stable law*. A distribution function F is stable if for each n there exist constants a_n and b_n, $a_n > 0$, such that, if $X_1, \ldots, X_n$ are independent and have distribution function F, then $a_n^{-1}(X_1 + \cdots + X_n) + b_n$ also has distribution function F.

28.10. Suppose that for all a, a', b, b' there exist a'', b'' (here a, a', a'' are all positive) such that $F(ax+b) * F(a'x+b') = F(a''x+b'')$. Show that F is stable.

28.11. Show that a stable law is infinitely divisible.

28.12. Show that the Poisson law, although infinitely divisible, is not stable.

28.13. Show that the normal and Cauchy laws are stable.

28.14. 28.10↑ Suppose that F has mean 0 and variance 1 and that the dependence of a'', b'' on a, a', b, b' is such that

$$F\left(\frac{x}{\sigma_1}\right) * F\left(\frac{x}{\sigma_2}\right) = F\left(\frac{x}{\sqrt{\sigma_1^2+\sigma_2^2}}\right).$$

Show that F is the standard normal distribution.

28.15. **(a)** Let Y_{nk} be independent random variables having the Poisson distribution with mean $cn^\alpha/|k|^{1+\alpha}$, where $c > 0$ and $0 < \alpha < 2$. Let $Z_n = n^{-1}\sum_{k=-n^2}^{n^2} kY_{nk}$ (omit $k = 0$ in the sum), and show that if c is properly chosen then the characteristic function of Z_n converges to $e^{-|t|^\alpha}$.

(b) Show for $0 < \alpha \leq 2$ that $e^{-|t|^\alpha}$ is the characteristic function of a symmetric stable distribution; it is called the *symmetric stable law of exponent* α. The case $\alpha = 2$ is the normal law, and $\alpha = 1$ is the Cauchy law.

SECTION 29. LIMIT THEOREMS IN R^k

If F_n and F are distribution functions on R^k, then F_n *converges weakly* to F, written $F_n \Rightarrow F$, if $\lim_n F_n(x) = F(x)$ for all continuity points x of F. The corresponding distributions μ_n and μ are in this case also said to converge weakly: $\mu_n \Rightarrow \mu$. If X_n and X are k-dimensional random vectors (possibly on different probability spaces), X_n *converges in distribution* to X, written $X_n \Rightarrow X$, if the corresponding distribution functions converge weakly. The definitions are thus exactly as for the line.

The Basic Theorems

The closure A^- of a set in R^k is the set of limits of sequences in A; the interior is $A^\circ = R^k - (R^k - A)^-$; and the boundary is $\partial A = A^- - A^\circ$. A Borel set A is a *μ-continuity set* if $\mu(\partial A) = 0$. The first theorem is the k-dimensional version of Theorem 25.8.

Theorem 29.1. *For probability measures μ_n and μ on $(R^k, \mathscr{R}^k)$, each of the following conditions is equivalent to the weak convergence of μ_n to μ:*

(i) $\lim_n \int f\,d\mu_n = \int f\,d\mu$ *for bounded continuous f;*
(ii) $\limsup_n \mu_n(C) \leq \mu(C)$ *for closed C;*
(iii) $\liminf_n \mu_n(G) \geq \mu(G)$ *for open G;*
(iv) $\lim_n \mu_n(A) = \mu(A)$ *for μ-continuity sets A.*

PROOF. It will first be shown that (i) through (iv) are all equivalent.

(i) *implies* (ii): Consider the distance $\operatorname{dist}(x, C) = \inf[|x - y|: y \in C]$ from x to C. It is continuous in x. Let

$$\varphi_j(t) = \begin{cases} 1 & \text{if } t \leq 0, \\ 1 - jt & \text{if } 0 \leq t \leq j^{-1}, \\ 0 & \text{if } j^{-1} \leq t. \end{cases}$$

Then $f_j(x) = \varphi_j(\text{dist}(x, C))$ is continuous and bounded by 1, and $f_j(x) \downarrow I_C(x)$ as $j \uparrow \infty$ because C is closed. If (i) holds, then $\limsup_n \mu_n(C) \leq \lim_n \int f_j \, d\mu_n = \int f_j \, d\mu$. As $j \uparrow \infty$, $\int f_j \, d\mu \downarrow \int I_C \, d\mu = \mu(C)$.

(ii) *is equivalent to* (iii). Take $C = R^k - G$.

(ii) *and* (iii) *imply* (iv): From (ii) and (iii) follows

$$\mu(A^\circ) \leq \liminf_n \mu_n(A^\circ) \leq \liminf_n \mu_n(A)$$

$$\leq \limsup_n \mu_n(A) \leq \limsup_n \mu_n(A^-) \leq \mu(A^-).$$

Clearly (iv) follows from this.

(iv) *implies* (i): Suppose that f is continuous and $|f(x)|$ is bounded by K. Given ϵ, choose reals $\alpha_0 < \alpha_1 < \cdots < \alpha_l$ so that $\alpha_0 < -K < K < \alpha_l$, $\alpha_i - \alpha_{i-1} < \epsilon$, and $\mu[x\colon f(x) = \alpha_i] = 0$. The last condition can be achieved because the sets $[x\colon f(x) = \alpha]$ are disjoint for different α. Put $A_i = [x\colon \alpha_{i-1} < f(x) \leq \alpha_i]$. Since f is continuous, $A_i^- \subset [x\colon \alpha_{i-1} \leq f(x) \leq \alpha_i]$ and $A_i^\circ \supset [x\colon \alpha_{i-1} < f(x) < \alpha_i]$. Therefore, $\partial A_i \subset [x\colon f(x) = \alpha_{i-1}] \cup [x\colon f(x) = \alpha_i]$, and therefore $\mu(\partial A_i) = 0$. Now $|\int f \, d\mu_n - \sum_{i=1}^l \alpha_i \mu_n(A_i)| \leq \epsilon$ and similarly for μ, and $\sum_{i=1}^l \alpha_i \mu_n(A_i) \to \sum_{i=1}^l \alpha_i \mu(A_i)$ because of (iv). Since ϵ was arbitrary, (i) follows.

It remains to prove these four conditions equivalent to weak convergence.

(iv) *implies* $\mu_n \Rightarrow \mu$: Consider the corresponding distribution functions. If $S_x = [y\colon y_i \leq x_i,\ i = 1, \ldots, k]$, then F is continuous at x if and only if $\mu(\partial S_x) = 0$; see the argument following (20.18). Therefore, if F is continuous at x, $F_n(x) = \mu_n(S_x) \to \mu(S_x) = F(x)$, and $F_n \Rightarrow F$.

$\mu_n \Rightarrow \mu$ *implies* (iii): Since only countably many parallel hyperplanes can have positive μ-measure, there is a dense set D of reals such that $\mu[x\colon x_i = d] = 0$ for $d \in D$ and $i = 1, \ldots, k$. Let $\mathscr{A}$ be the class of rectangles $A = [x\colon a_i < x_i \leq b_i,\ i = 1, \ldots, k]$ for which the a_i and the b_i all lie in D. All 2^k vertices of such a rectangle are continuity points of F, and so $F_n \Rightarrow F$ implies (see (12.12)) that $\mu_n(A) = \Delta_A F_n \to \Delta_A F = \mu(A)$. It follows by the inclusion–exclusion formula that $\mu_n(B) \to \mu(B)$ for finite unions B of elements of $\mathscr{A}$. Since D is dense on the line, an open set G in R^k is a countable union of sets A_m in $\mathscr{A}$. But $\mu(\bigcup_{m \leq M} A_m) = \lim_n \mu_n(\bigcup_{m \leq M} A_m) \leq \liminf_n \mu_n(G)$. Letting $M \to \infty$ gives (iii). ■

Theorem 29.2. *Suppose that* $h\colon R^k \to R^j$ *is measurable and that the set* D_h *of its discontinuities is measurable.*[†] *If* $\mu_n \Rightarrow \mu$ *in* R^k *and* $\mu(D_h) = 0$, *then* $\mu_n h^{-1} \Rightarrow \mu h^{-1}$ *in* R^j.

[†]The argument in the footnote on p. 334 shows that in fact $D_h \in \mathscr{R}^k$ always holds.

PROOF. Let C be a closed set in R^j. The closure $(h^{-1}C)^-$ in R^k satisfies $(h^{-1}C)^- \subset D_h \cup h^{-1}C$. If $\mu_n \Rightarrow \mu$, then part (ii) of Theorem 29.1 gives

$$\limsup_n \mu_n h^{-1}(C) \le \limsup_n \mu_n\bigl((h^{-1}C)^-\bigr) \le \mu\bigl((h^{-1}C)^-\bigr)$$

$$\le \mu(D_h) + \mu\bigl(h^{-1}C\bigr).$$

Using (ii) again gives $\mu_n h^{-1} \Rightarrow \mu h^{-1}$ if $\mu(D_h) = 0$. ■

Theorem 29.2 is the k-dimensional version of the *mapping theorem*—Theorem 25.7. The two proofs just given provide in the case $k = 1$ a second approach to the theory of Section 25, which there was based on Skorohod's theorem (Theorem 25.6). Skorohod's theorem does extend to R^k, but the proof is harder.†

Theorems 29.1 and 29.2 can of course be stated in terms of random vectors. For example, $X_n \Rightarrow X$ if and only if $P[X \in G] \le \liminf_n P[X_n \in G]$ for all open sets G.

A sequence $\{\mu_n\}$ of probability measures on $(R^k, \mathscr{R}^k)$ is *tight* if for every ϵ there is a bounded rectangle A such that $\mu_n(A) > 1 - \epsilon$ for all n.

Theorem 29.3. *If $\{\mu_n\}$ is a tight sequence of probability measures, there is a subsequence $\{\mu_{n_i}\}$ and a probability measure μ such that $\mu_{n_i} \Rightarrow \mu$ as $i \to \infty$.*

PROOF. Take $S_x = [y\colon y_j \le x_j,\ j \le k]$ and $F_n(x) = \mu_n(S_x)$. The proof of Helly's theorem (Theorem 25.9) carries over: For points x and y in R^k, interpret $x \le y$ as meaning $x_u \le y_u$, $u = 1, \ldots, k$, and $x < y$ as meaning $x_u < y_u$, $u = 1, \ldots, k$. Consider rational points r—points whose coordinates are all rational—and by the diagonal method [A14] choose a sequence $\{n_i\}$ along which $\lim_i F_{n_i}(r) = G(r)$ exists for each such r. As before, define $F(x) = \inf[G(r)\colon x < r]$. Although F is clearly nondecreasing in each variable, a further argument is required to prove $\Delta_A F \ge 0$ (see (12.12)).

Given ϵ and a rectangle $A = (a_1, b_1] \times \cdots \times (a_k, b_k]$, choose a δ such that if $z = (\delta, \ldots, \delta)$, then for each of the 2^k vertices x of A, $x < r < x + z$ implies $|F(x) - G(r)| < \epsilon/2^k$. Now choose rational points r and s such that $a < r < a + z$ and $b < s < b + z$. If $B = (r_1, s_1] \times \cdots \times (r_k, s_k]$, then $|\Delta_A F - \Delta_B G| < \epsilon$. Since $\Delta_B G = \lim_i \Delta_B F_{n_i} \ge 0$ and ϵ is arbitrary, it follows that $\Delta_A F \ge 0$.

With the present interpretation of the symbols, the proof of Theorem 25.9 shows that F is continuous from above and $\lim_i F_{n_i}(x) = F(x)$ for continuity points x of F.

†The approach of this section carries over to general metric spaces; for this theory and its applications, see BILLINGSLEY$_1$ and BILLINGSLEY$_2$. Since Skorohod's theorem is no easier in R^k than in the general metric space, it is not treated here.

By Theorem 12.5, there is a measure μ on $(R^k, \mathscr{R}^k)$ such that $\mu(A) = \Delta_A F$ for rectangles A. By tightness, there is for given ϵ a t such that $\mu_n[y: -t < y_j \le t,\ j \le k] > 1 - \epsilon$ for all n. Suppose that all coordinates of x exceed t: If $r > x$, then $F_n(r) > 1 - \epsilon$ and hence (r rational) $G(r) \ge 1 - \epsilon$, so that $F(x) \ge 1 - \epsilon$. Suppose, on the other hand, that some coordinate of x is less than $-t$: Choose a rational r such that $x < r$ and some coordinate of r is less than $-t$; then $F_n(r) < \epsilon$, hence $G(r) \le \epsilon$, and so $F(x) \le \epsilon$. Therefore, for every ϵ there is a t such that

$$F(x)\begin{cases} \ge 1 - \epsilon & \text{if } x_j > t \text{ for all } j, \\ \le \epsilon & \text{if } x_j < -t \text{ for some } j. \end{cases} \tag{29.1}$$

If $B_s = [y: -s < y_j \le x_j,\ j \le k]$, then $\mu(S_x) = \lim_s \mu(B_s) = \lim_s \Delta_{B_s} F$. Of the 2^k terms in the sum $\Delta_{B_s} F$, all but $F(x)$ go to 0 ($s \to \infty$) because of the second part of (29.1). Thus $\mu(S_x) = F(x)$.[†] Because of the other part of (29.1), μ is a probability measure. Therefore, $F_n \Rightarrow F$ and $\mu_n \Rightarrow \mu$. ■

Obviously Theorem 29.3 implies that tightness is a sufficient condition that each subsequence of $\{\mu_n\}$ contain a further subsequence converging weakly to some probability measure. (An easy modification of the proof of Theorem 25.10 shows that tightness is necessary for this as well.) And clearly the corollary to Theorem 25.10 now goes through:

Corollary. *If $\{\mu_n\}$ is a tight sequence of probability measures, and if each subsequence that converges weakly at all converges weakly to the probability measure μ, then $\mu_n \Rightarrow \mu$.*

Characteristic Functions

Consider a random vector $X = (X_1, \ldots, X_k)$ and its distribution μ in R^k. Let $t \cdot x = \sum_{u=1}^k t_u x_u$ denote inner product. The characteristic function of X and of μ is defined over R^k by

$$\varphi(t) = \int_{R^k} e^{it \cdot x} \mu(dx) = E[e^{it \cdot X}]. \tag{29.2}$$

To a great extent its properties parallel those of the one-dimensional characteristic function and can be deduced by parallel arguments.

[†]This requires proof because there exist (Problem 12.10) functions F' other than F for which $\mu(A) = \Delta_A F'$ holds for all rectangles A.

The inversion formula (26.16) takes this form: For a bounded rectangle $A = [x: a_u < x_u \le b_u,\ u \le k]$ such that $\mu(\partial A) = 0$,

$$\mu(A) = \lim_{T\to\infty} \frac{1}{(2\pi)^k} \int_{B_T} \prod_{u=1}^{k} \frac{e^{-it_u a_u} - e^{-it_u b_u}}{it_u} \varphi(t)\, dt, \tag{29.3}$$

where $B_T = [t \in R^k: |t_u| \le T,\ u \le k]$ and dt is short for $dt_1 \cdots dt_k$. To prove it, replace $\varphi(t)$ by the middle term in (29.2) and reverse the integrals as in (26.17): The integral in (29.3) is

$$I_T = \frac{1}{(2\pi)^k} \int_{R^k} \left[\int_{B_T} \prod_{u=1}^{k} \frac{e^{-it_u a_u} - e^{-it_u b_u}}{it_u} e^{it_u x_u}\, dt \right] \mu(dx).$$

The inner integral may be evaluated by Fubini's theorem in R^k, which gives

$$I_T = \int_{R^k} \prod_{u=1}^{k} \left[\frac{\operatorname{sgn}(x_u - a_u)}{\pi} S(T\cdot|x_u - a_u|) - \frac{\operatorname{sgn}(x_u - b_u)}{\pi} S(T\cdot|x_u - b_u|) \right] \mu(dx).$$

Since the integrand converges to $\prod_{u=1}^{k} \psi_{a_u, b_u}(x_u)$ (see (26.18)), (29.3) follows as in the case $k = 1$.

The proof that weak convergence implies (iii) in Theorem 29.1 shows that for probability measures μ and ν on R^k there exists a dense set D of reals such that $\mu(\partial A) = \nu(\partial A) = 0$ for all rectangles A whose vertices have coordinates in D. If $\mu(A) = \nu(A)$ for such rectangles, then μ and ν are identical by Theorem 3.3.

Thus *the characteristic function φ uniquely determines the probability measure μ*. Further properties of the characteristic function can be derived from the one-dimensional case by means of the following device of Cramér and Wold. For $t \in R^k$, define h_t: $R^k \to R^1$ by $h_t(x) = t \cdot x$. For real α, $[x: t \cdot x \le \alpha]$ is a half space, and its μ-measure is

$$\mu[x: t \cdot x \le \alpha] = \mu h_t^{-1}(-\infty, \alpha]. \tag{29.4}$$

By change of variable, the characteristic function of μh_t^{-1} is

$$\begin{aligned} \int_{R^1} e^{isy} \mu h_t^{-1}(dy) &= \int_{R^k} e^{is(t\cdot x)} \mu(dx) \\ &= \varphi(st_1, \ldots, st_k), \qquad s \in R^1. \end{aligned} \tag{29.5}$$

To know the μ-measure of every half space is (by (29.4)) to know each μh_t^{-1} and hence is (by (29.5) for $s = 1$) to know $\varphi(t)$ for every t; and to know the

characteristic function φ of μ is to know μ. Thus μ *is uniquely determined by the values it gives to the half spaces*. This result, very simple in its statement, seems to require Fourier methods—no elementary proof is known.

If $\mu_n \Rightarrow \mu$ for probability measures on R^k, then $\varphi_n(t) \to \varphi(t)$ for the corresponding characteristic functions by Theorem 29.1. But suppose that the characteristic functions converge pointwise. It follows by (29.5) that for each t the characteristic function of $\mu_n h_t^{-1}$ converges pointwise on the line to the characteristic function of μh_t^{-1}; by the continuity theorem for characteristic functions on the line then, $\mu_n h_t^{-1} \Rightarrow \mu h_t^{-1}$. Take the uth component of t to be 1 and the others 0; then the $\mu_n h_t^{-1}$ are the marginals for the uth coordinate. Since $\{\mu_n h_t^{-1}\}$ is weakly convergent, there is a bounded interval $(a_u, b_u]$ such that $\mu_n[x \in R^k\colon a_u < x_u \le b_u] = \mu_n h_t^{-1}(a_u, b_u] > 1 - \epsilon/k$ for all n. But then $\mu_n(A) > 1 - \epsilon$ for the bounded rectangle $A = [x\colon a_u < x_u \le b_u, u = 1, \ldots, k]$. The sequence $\{\mu_n\}$ is therefore tight. If a subsequence $\{\mu_{n_i}\}$ converges weakly to ν, then $\varphi_{n_i}(t)$ converges to the characteristic function of ν, which is therefore $\varphi(t)$. By uniqueness, $\nu = \mu$, so that $\mu_{n_i} \Rightarrow \mu$. By the corollary to Theorem 29.3, $\mu_n \Rightarrow \mu$. This proves the continuity theorem for k-dimensional characteristic functions: $\mu_n \Rightarrow \mu$ *if and only if* $\varphi_n(t) \to \varphi(t)$ *for all* t.

The Cramér–Wold idea leads also to the following result, by means of which certain limit theorems can be reduced in a routine way to the one-dimensional case.

Theorem 29.4. *For random vectors* $X_n = (X_{n1}, \ldots, X_{nk})$ *and* $Y = (Y_1, \ldots, Y_k)$, *a necessary and sufficient condition for* $X_n \Rightarrow Y$ *is that* $\sum_{u=1}^k t_u X_{nu} \Rightarrow \sum_{u=1}^k t_u Y_u$ *for each* $(t_1, \ldots, t_k)$ *in* R^k.

PROOF. The necessity follows from a consideration of the continuous mapping h_t above—use Theorem 29.2. As for sufficiency, the condition implies by the continuity theorem for one-dimensional characteristic functions that for each $(t_1, \ldots, t_k)$

$$E\left[e^{is\sum_{u=1}^k t_u X_{nu}}\right] \to E\left[e^{is\sum_{u=1}^k t_u Y_u}\right]$$

for all real s. Taking $s = 1$ shows that the characteristic function of X_n converges pointwise to that of Y. ■

Normal Distributions in R^k

By Theorem 20.4 there is (on some probability space) a random vector $X = (X_1, \ldots, X_k)$ with independent components each having the standard normal distribution. Since each X_u has density $e^{-x^2/2}/\sqrt{2\pi}$, X has density

(see (20.25))

$$f(x) = \frac{1}{(2\pi)^{k/2}} e^{-|x|^2/2}, \tag{29.6}$$

where $|x|^2 = \sum_{u=1}^{k} x_u^2$ denotes Euclidean norm. This distribution plays the role of the standard normal distribution in R^k. Its characteristic function is

$$E\left[\prod_{u=1}^{k} e^{it_u X_u}\right] = \prod_{u=1}^{k} e^{-t_u^2/2} = e^{-|t|^2/2}. \tag{29.7}$$

Let $A = [a_{uv}]$ be a $k \times k$ matrix, and put $Y = AX$, where X is viewed as a column vector. Since $E[X_\alpha X_\beta] = \delta_{\alpha\beta}$, the matrix $\Sigma = [\sigma_{uv}]$ of the covariances of Y has entries $\sigma_{uv} = E[Y_u Y_v] = \sum_{\alpha=1}^{k} a_{u\alpha} a_{v\alpha}$. Thus $\Sigma = AA'$, where the prime denotes transpose. The matrix Σ is symmetric and nonnegative definite: $\sum_{uv} \sigma_{uv} x_u x_v = |A'x|^2 \geq 0$. View t also as a column vector with transpose t', and note that $t \cdot x = t'x$. The characteristic function of AX is thus

$$E[e^{it'(AX)}] = E[e^{i(A't)'X}] = e^{-|A't|^2/2} = e^{-t'\Sigma t/2}. \tag{29.8}$$

Define a centered *normal distribution* as any probability measure whose characteristic function has this form for some symmetric nonnegative definite Σ.

If Σ is symmetric and nonnegative definite, then for an appropriate orthogonal matrix U, $U'\Sigma U = D$ is a diagonal matrix whose diagonal elements are the eigenvalues of Σ and hence are nonnegative. If D_0 is the diagonal matrix whose elements are the square roots of those of D, and if $A = UD_0$, then $\Sigma = AA'$. Thus for every nonnegative definite Σ there exists a centered normal distribution (namely the distribution of AX) with covariance matrix Σ and characteristic function $\exp(-\frac{1}{2}t'\Sigma t)$.

If Σ is nonsingular, so is the A just constructed. Since X has density (29.6), $Y = AX$ has, by the Jacobian transformation formula (20.20), density $f(A^{-1}x)|\det A^{-1}|$. From $\Sigma = AA'$ follows $|\det A^{-1}| = (\det \Sigma)^{-1/2}$. Moreover, $\Sigma^{-1} = (A')^{-1}A^{-1}$, so that $|A^{-1}x|^2 = x'\Sigma^{-1}x$. Thus the normal distribution has density $(2\pi)^{k/2}(\det \Sigma)^{-1/2} \exp(-\frac{1}{2}x'\Sigma^{-1}x)$ if Σ is nonsingular. If Σ is singular, the A constructed above must be singular as well, so that AX is confined to some hyperplane of dimension $k-1$ and the distribution can have no density.

By (29.8) and the uniqueness theorem for characteristic functions in R^k, *a centered normal distribution is completely determined by its covariance matrix.* Suppose the off-diagonal elements of Σ are 0, and let A be the diagonal matrix with the $\sigma_{ii}^{1/2}$ along the diagonal. Then $\Sigma = AA'$, and if X has the standard normal distribution, the components X_i are independent and hence so are the components $\sigma_{ii}^{1/2} X_i$ of AX. Therefore, *the components of a*

normally distributed random vector are independent if and only if they are uncorrelated.

If M is a $j \times k$ matrix and Y has in R^k the centered normal distribution with covariance matrix Σ, then MY has in R^j the characteristic function $\exp(-\frac{1}{2}(M't)'\Sigma(M't)) = \exp(-\frac{1}{2}t'(M\Sigma M')t)$ $(t \in R^j)$. Hence MY has the centered normal distribution in R^j with covariance matrix $M\Sigma M'$. Thus *a linear transformation of a normal distribution is itself normal.*

These normal distributions are special in that all the first moments vanish. The general normal distribution is a translation of one of these centered distributions. It is completely determined by its means and covariances.

The Central Limit Theorem

Let $X_n = (X_{n1}, \ldots, X_{nk})$ be independent random vectors all having the same distribution. Suppose that $E[X_{nu}^2] < \infty$; let the vector of means be $c = (c_1, \ldots, c_k)$, where $c_u = E[X_{nu}]$, and let the covariance matrix be $\Sigma = [\sigma_{uv}]$, where $\sigma_{uv} = E[(X_{nu} - c_u)(X_{nv} - c_v)]$. Put $S_n = X_1 + \cdots + X_n$.

Theorem 29.5. *Under these assumptions, the distribution of the random vector* $(S_n - nc)/\sqrt{n}$ *converges weakly to the centered normal distribution with covariance matrix* Σ.

PROOF. Let $Y = (Y_1, \ldots, Y_k)$ be a normally distributed random vector with 0 means and covariance matrix Σ. For given $t = (t_1, \ldots, t_k)$, let $Z_n = \sum_{u=1}^k t_u(X_{nu} - c_u)$ and $Z = \sum_{u=1}^k t_u Y_u$. By Theorem 29.4, it suffices to prove that $n^{-1/2}\sum_{j=1}^n Z_j \Rightarrow Z$ (for arbitrary t). But this is an instant consequence of the Lindeberg–Lévy theorem (Theorem 27.1). ■

PROBLEMS

29.1. A real function f on R^k is everywhere upper semicontinuous (see Problem 13.8) if for each x and ϵ there is a δ such that $|x - y| < \delta$ implies that $f(y) < f(x) + \epsilon$; f is lower semicontinuous if $-f$ is upper semicontinuous.

(a) Use condition (iii) of Theorem 29.1, Fatou's lemma, and (21.9) to show that, if $\mu_n \Rightarrow \mu$ and f is bounded and lower semicontinuous, then

$$\liminf_n \int f\,d\mu_n \geq \int f\,d\mu. \tag{29.9}$$

(b) Show that, if (29.9) holds for all bounded, lower semicontinuous functions f, then $\mu_n \Rightarrow \mu$.

(c) Prove the analogous results for upper semicontinuous functions.

29.2. (a) Show for probability measures on the line that $\mu_n \times \nu_n \Rightarrow \mu \times \nu$ if and only if $\mu_n \Rightarrow \mu$ and $\nu_n \Rightarrow \nu$.

(b) Suppose that X_n and Y_n are independent and that X and Y are independent. Show that, if $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$, then $(X_n, Y_n) \Rightarrow (X, Y)$ and hence that $X_n + Y_n \Rightarrow X + Y$.

(c) Show that part (b) fails without independence.

(d) If $F_n \Rightarrow F$ and $G_n \Rightarrow G$, then $F_n * G_n \Rightarrow F * G$. Prove this by part (b) and also by characteristic functions.

29.3. (a) Show that $\{\mu_n\}$ is tight if and only if for each ϵ there is a compact set K such that $\mu_n(K) > 1 - \epsilon$ for all n.

(b) Show that $\{\mu_n\}$ is tight if and only if each of the k sequences of marginal distributions is tight on the line.

29.4. Assume of (X_n, Y_n) that $X_n \Rightarrow X$ and $Y_n \Rightarrow c$. Show that $(X_n, Y_n) \Rightarrow (X, c)$. This is an example of Problem 29.2(b) where X_n and Y_n need not be assumed independent.

29.5. Prove analogues for R^k of the corollaries to Theorem 26.3.

29.6. Suppose that $f(X)$ and $g(Y)$ are uncorrelated for all bounded continuous f and g. Show that X and Y are independent. *Hint*: Use characteristic functions.

29.7. 20.16↑ Suppose that the random vector X has a centered k-dimensional normal distribution whose covariance matrix has 1 as an eigenvalue of multiplicity r and 0 as an eigenvalue of multiplicity $k - r$. Show that $|X|^2$ has the chi-squared distribution with r degrees of freedom.

29.8. ↑ *Multinomial sampling.* Let $p_1, \ldots, p_k$ be positive and add to 1, and let $Z_1, Z_2, \ldots$ be independent k-dimensional random vectors such that Z_n has with probability p_i a 1 in the ith component and 0's elsewhere. Then $f_n = (f_{n1}, \ldots, f_{nk}) = \sum_{m=1}^{n} Z_m$ is the frequency count for a sample of size n from a multinomial population with cell probabilities p_i. Put $X_{ni} = (f_{ni} - np_i)/\sqrt{np_i}$ and $X_n = (X_{n1}, \ldots, X_{nk})$.

(a) Show that X_n has mean values 0 and covariances $\sigma_{ij} = (\delta_{ij} p_j - p_i p_j)/\sqrt{p_i p_j}$.

(b) Show that the chi squared statistic $\sum_{i=1}^{k} (f_{ni} - np_i)^2/np_i$ has asymptotically the chi-squared distribution with $k - 1$ degrees of freedom.

29.9. 20.26↑ *A theorem of Poincaré.* (a) Suppose that $X_n = (X_{n1}, \ldots, X_{nn})$ is uniformly distributed over the surface of a sphere of radius $\sqrt{n}$ in R^n. Fix t, and show that $X_{n1}, \ldots, X_{nt}$ are in the limit independent, each with the standard normal distribution. *Hint:* If the components of $Y_n = (Y_{n1}, \ldots, Y_{nn})$ are independent, each with the standard normal distribution, then X_n has the same distribution as $\sqrt{n}\, Y_n/|Y_n|$.

(b) Suppose that the distribution of $X_n = (X_{n1}, \ldots, X_{nn})$ is spherically symmetric in the sense that $X_n/|X_n|$ is uniformly distributed over the unit sphere. Assume that $|X_n|^2/n \Rightarrow 1$, and show that $X_{n1}, \ldots, X_{nt}$ are asymptotically independent and normal.

29.10. Let $X_n = (X_{n1}, \ldots, X_{nk})$, $n = 1, 2, \ldots$, be random vectors satisfying the mixing condition (27.19) with $\alpha_n = O(n^{-5})$. Suppose that the sequence is stationary (the distribution of $(X_n, \ldots, X_{n+j})$ is the same for all n), that $E[X_{nu}] = 0$, and that the X_{nu} are uniformly bounded. Show that if $S_n = X_1 + \cdots + X_n$, then $S_n/\sqrt{n}$ has in the limit the centered normal distribution with covariances

$$E[X_{1u}X_{1v}] + \sum_{j=1}^{\infty} E[X_{1u}X_{1+j,v}] + \sum_{j=1}^{\infty} E[X_{1+j,u}X_{1v}].$$

Hint: Use the Cramér–Wold device.

29.11. ↑ As in Example 27.6, let $\{Y_n\}$ be a Markov chain with finite state space $S = \{1, \ldots, s\}$, say. Suppose the transition probabilities p_{uv} are all positive and the initial probabilities p_u are the stationary ones. Let f_{nu} be the number of i for which $1 \le i \le n$ and $Y_i = u$. Show that the normalized frequency count

$$n^{-1/2}(f_{n1} - np_1, \ldots, f_{nk} - np_k)$$

has in the limit the centered normal distribution with covariances

$$\delta_{uv} - p_u p_v + \sum_{j=1}^{\infty} \left(p_{uv}^{(j)} - p_u p_v\right) + \sum_{j=1}^{\infty} \left(p_{vu}^{(j)} - p_v p_u\right).$$

29.12. Assume that

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

is positive definite, invert it explicitly, and show that the corresponding two-dimensional normal density is

$$f(x_1, x_2) = \frac{1}{2\pi D^{1/2}} \exp\left[-\frac{1}{2D}\left(\sigma_{22}x_1^2 - 2\sigma_{12}x_1x_2 + \sigma_{11}x_2^2\right)\right], \tag{29.10}$$

where $D = \sigma_{11}\sigma_{22} - \sigma_{12}^2$.

29.13. Suppose that Z has the standard normal distribution in R^1. Let μ be the mixture with equal weights of the distributions of (Z, Z) and $(Z, -Z)$, and let (X, Y) have distribution μ. Prove:

(a) Although each of X and Y is normal, they are not jointly normal.

(b) Although X and Y are uncorrelated, they are not independent.

SECTION 30. THE METHOD OF MOMENTS*

The Moment Problem

For some distributions the characteristic function is intractable but moments can nonetheless be calculated. In these cases it is sometimes possible to prove weak convergence of the distributions by establishing that the moments converge. This approach requires conditions under which a distribution is uniquely determined by its moments, and this is for the same reason that the continuity theorem for characteristic functions requires for its proof the uniqueness theorem.

Theorem 30.1. *Let μ be a probability measure on the line having finite moments $\alpha_k = \int_{-\infty}^{\infty} x^k \mu(dx)$ of all orders. If the power series $\sum_k \alpha_k r^k/k!$ has a positive radius of convergence, then μ is the only probability measure with the moments $\alpha_1, \alpha_2, \ldots$.*

PROOF. Let $\beta_k = \int_{-\infty}^{\infty} |x|^k \mu(dx)$ be the absolute moments. The first step is to show that

$$\frac{\beta_k r^k}{k!} \to 0, \qquad k \to \infty, \tag{30.1}$$

for some positive r. By hypothesis there exists an s, $0 < s < 1$, such that $\alpha_k s^k/k! \to 0$. Choose $0 < r < s$; then $2kr^{2k-1} < s^{2k}$ for large k. Since $|x|^{2k-1} \leq 1 + |x|^{2k}$,

$$\frac{\beta_{2k-1} r^{2k-1}}{(2k-1)!} \leq \frac{r^{2k-1}}{(2k-1)!} + \frac{\beta_{2k} s^{2k}}{(2k)!}$$

for large k. Hence (30.1) holds as k goes to infinity through odd values; since $\beta_k = \alpha_k$ for k even, (30.1) follows.

By (26.4),

$$\left| e^{itx} \left(e^{ihx} - \sum_{k=0}^{n} \frac{(ihx)^k}{k!} \right) \right| \leq \frac{|hx|^{n+1}}{(n+1)!},$$

and therefore the characteristic function φ of μ satisfies

$$\left| \varphi(t+h) - \sum_{k=0}^{n} \frac{h^k}{k!} \int_{-\infty}^{\infty} (ix)^k e^{itx} \mu(dx) \right| \leq \frac{|h|^{n+1} \beta_{n+1}}{(n+1)!}.$$

*This section may be omitted.

By (26.10), the integral here is $\varphi^{(k)}(t)$. By (30.1),

$$\varphi(t+h) = \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(t)}{k!} h^k, \qquad |h| \le r. \tag{30.2}$$

If ν is another probability measure with moments α_k and characteristic function $\psi(t)$, the same argument gives

$$\psi(t+h) = \sum_{k=0}^{\infty} \frac{\psi^{(k)}(t)}{k!} h^k, \qquad |h| \le r. \tag{30.3}$$

Take $t = 0$; since $\varphi^{(k)}(0) = i^k \alpha_k = \psi^{(k)}(0)$ (see (26.9)), φ and ψ agree in $(-r, r)$ and hence have identical derivatives there. Taking $t = r - \epsilon$ and $t = -r + \epsilon$ in (30.2) and (30.3) shows that φ and ψ also agree in $(-2r + \epsilon, 2r - \epsilon)$ and hence in $(-2r, 2r)$. But then they must by the same argument agree in $(-3r, 3r)$ as well, and so on.† Thus φ and ψ coincide, and by the uniqueness theorem for characteristic functions, so do μ and ν. ■

A probability measure satisfying the conclusion of the theorem is said to be *determined by its moments*.

Example 30.1. For the standard normal distribution, $|\alpha_k| \le k!$, and so the theorem implies that it is determined by its moments. ■

But not all measures are determined by their moments:

Example 30.2. If N has the standard normal density, then e^N has the log-normal density

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \dfrac{1}{x} e^{-(\log x)^2/2} & \text{if } x > 0, \\ 0 & \text{if } x \le 0. \end{cases}$$

Put $g(x) = f(x)(1 + \sin(2\pi \log x))$. If

$$\int_0^{\infty} x^k f(x) \sin(2\pi \log x)\, dx = 0, \qquad k = 0, 1, 2, \ldots,$$

then g, which is nonnegative, will be a probability density and will have the same moments as f. But a change of variable $\log x = s + k$ reduces the

†This process is a version of analytic continuation.

integral above to

$$\frac{1}{\sqrt{2\pi}} e^{k^2/2} \int_{-\infty}^{\infty} e^{-s^2/2} \sin 2\pi s \, ds,$$

which vanishes because the integrand is odd. ■

Theorem 30.2. *Suppose that the distribution of X is determined by its moments, that the X_n have moments of all orders, and that $\lim_n E[X_n^r] = E[X^r]$ for $r = 1, 2, \ldots$. Then $X_n \Rightarrow X$.*

PROOF. Let μ_n and μ be the distributions of X_n and X. Since $E[X_n^2]$ converges, it is bounded, say by K. By Markov's inequality, $P[|X_n| \geq x] \leq K/x^2$, which implies that the sequence $\{\mu_n\}$ is tight.

Suppose that $\mu_{n_k} \Rightarrow \nu$, and let Y be a random variable with distribution ν. If u is an even integer exceeding r, the convergence and hence boundedness of $E[X_n^u]$ implies that $E[X_{n_k}^r] \to E[Y^r]$, by the corollary to Theorem 25.12. By the hypothesis, then, $E[Y^r] = E[X^r]$—that is, ν and μ have the same moments. Since μ is by hypothesis determined by its moments, ν must be the same as μ, and so $\mu_{n_k} \Rightarrow \mu$. The conclusion now follows by the corollary to Theorem 25.10. ■

Convergence to the log-normal distribution cannot be proved by establishing convergence of moments (take X to have density f and the X_n to have density g in Example 30.2). Because of Example 30.1, however, this approach will work for a normal limit.

Moment Generating Functions

Suppose that μ has a moment generating function $M(s)$ for $s \in [-s_0, s_0]$, $s_0 > 0$. By (21.22), the hypothesis of Theorem 30.1 is satisfied, and so μ is determined by its moments, which are in turn determined by $M(s)$ via (21.23). Thus *μ is determined by $M(s)$ if it exists in a neighborhood of 0.*[†] The version of this for one-sided transforms was proved in Section 22—see Theorem 22.2.

Suppose that μ_n and μ have moment generating functions in a common interval $[-s_0, s_0]$, $s_0 > 0$, and suppose that $M_n(s) \to M(s)$ in this interval. Since $\mu_n[(-a, a)^c] \leq e^{-s_0 a}(M_n(-s_0) + M_n(s_0))$, it follows easily that $\{\mu_n\}$ is tight. Since $M(s)$ determines μ, the usual argument now gives $\mu_n \Rightarrow \mu$.

[†]For another proof, see Problem 26.7. The present proof does not require the idea of analyticity.

Central Limit Theorem by Moments

To understand the application of the method of moments, consider once again a sum $S_n = X_{n1} + \cdots + X_{nk_n}$, where $X_{n1}, \ldots, X_{nk_n}$ are independent and

$$E[X_{nk}] = 0, \qquad E[X_{nk}^2] = \sigma_{nk}^2, \qquad s_n^2 = \sum_{k=1}^{k_n} \sigma_{nk}^2. \tag{30.4}$$

Suppose further that for each n there is an M_n such that $|X_{nk}| \le M_n$, $k = 1, \ldots, k_n$, with probability 1. Finally, suppose that

$$\frac{M_n}{s_n} \to 0. \tag{30.5}$$

All moments exist, and[†]

$$S_n^r = \sum_{u=1}^{r} \sum{}' \frac{r!}{r_1! \cdots r_u!} \frac{1}{u!} \sum{}'' X_{ni_1}^{r_1} \cdots X_{ni_u}^{r_u}, \tag{30.6}$$

where $\sum'$ extends over the u-tuples $(r_1, \ldots, r_u)$ of positive integers satisfying $r_1 + \cdots + r_u = r$ and $\sum''$ extends over the u-tuples $(i_1, \ldots, i_u)$ of distinct integers in the range $1 \le i_\alpha \le k_n$.

By independence, then,

$$E\left[\left(\frac{S_n}{s_n}\right)^r\right] = \sum_{u=1}^{r} \sum{}' \frac{r!}{r_1! \cdots r_u!} \frac{1}{u!} A_n(r_1, \ldots, r_u), \tag{30.7}$$

where

$$A_n(r_1, \ldots, r_u) = \sum{}'' \frac{1}{s_n^r} E[X_{ni_1}^{r_1}] \cdots E[X_{ni_u}^{r_u}], \tag{30.8}$$

and $\sum'$ and $\sum''$ have the same ranges as before. To prove that (30.7) converges to the rth moment of the standard normal distribution, it suffices to show that

$$\lim_n A_n(r_1, \ldots, r_u) = \begin{cases} 1 & \text{if } r_1 = \cdots = r_u = 2, \\ 0 & \text{otherwise} \end{cases}. \tag{30.9}$$

Indeed, if r is even, all terms in (30.7) will then go to 0 except the one for which $u = r/2$ and $r_\alpha \equiv 2$, which will go to $r!/(r_1! \cdots r_u! u!) = 1 \times 3 \times 5 \times \cdots \times (r-1)$. And if r is odd, the terms will go to 0 without exception.

[†]To deduce this from the multinomial formula, restrict the inner sum to u-tuples satisfying $1 \le i_1 < \cdots < i_u \le k_n$ and compensate by striking out the $1/u!$.

If $r_\alpha = 1$ for some α, then (30.9) holds because by (30.4) each summand in (30.8) vanishes. Suppose that $r_\alpha \geq 2$ for each α and $r_\alpha > 2$ for some α. Then $r > 2u$, and since $|E[X_{ni}^{r_\alpha}]| \leq M_n^{(r_\alpha - 2)}\sigma_{ni}^2$, it follows that $A_n(r_1, \ldots, r_u) \leq (M_n/s_n)^{r-2u} A_n(2, \ldots, 2)$. But this goes to 0 because (30.5) holds and because $A_n(2, \ldots, 2)$ is bounded by 1 (it increases to 1 if the sum in (30.8) is enlarged to include all the u-tuples $(i_1, \ldots, i_u)$).

It remains only to check (30.9) for $r_1 = \cdots = r_u = 2$. As just noted, $A_n(2, \ldots, 2)$ is at most 1, and it differs from 1 by $\sum s_n^{-2u}\sigma_{ni_u}^2$, the sum extending over the $(i_1, \ldots, i_u)$ with at least one repeated index. Since $\sigma_{ni}^2 \leq M_n^2$, the terms for example with $i_u = i_{u-1}$ sum to at most $M_n^2 s_n^{-2u} \sum \sigma_{ni_1}^2 \cdots \sigma_{ni_{u-1}}^2 \leq M_n^2 s_n^{-2}$. Thus $1 - A_n(2, \ldots, 2) \leq u^2 M_n^2 s_n^{-2} \to 0$.

This proves that the moments (30.7) converge to those of the normal distribution and hence that $S_n/s_n \Rightarrow N$.

Application to Sampling Theory

Suppose that n numbers

$$x_{n1}, x_{n2}, \ldots, x_{nn},$$

not necessarily distinct, are associated with the elements of a population of size n. Suppose that these numbers are normalized by the requirement

$$\text{(30.10)} \qquad \sum_{h=1}^{n} x_{nh} = 0, \qquad \sum_{h=1}^{n} x_{nh}^2 = 1, \qquad M_n = \max_{h \leq n} |x_{nh}|.$$

An ordered sample $X_{n1}, \ldots, X_{nk_n}$ is taken, where the sampling is without replacement. By (30.10), $E[X_{nk}] = 0$ and $E[X_{nk}^2] = 1/n$. Let $s_n^2 = k_n/n$ be the fraction of the population sampled. If the X_{nk} were independent, which they are not, $S_n = X_{n1} + \cdots + X_{nk_n}$ would have variance s_n^2. If k_n is small in comparison with n, the effects of dependence should be small. It will be shown that $S_n/s_n \Rightarrow N$ if

$$\text{(30.11)} \qquad s_n^2 = \frac{k_n}{n} \to 0, \qquad \frac{M_n}{s_n} \to 0, \qquad k_n \to \infty.$$

Since $M_n^2 \geq n^{-1}$ by (30.10), the second condition here in fact implies the third.

The moments again have the form (30.7), but this time $E[X_{ni_1}^{r_1} \cdots X_{ni_u}^{r_u}]$ cannot be factored as in (30.8). On the other hand, this expected value is by symmetry the same for each of the $(k_n)_u = k_n(k_n - 1) \cdots (k_n - u + 1)$ choices of the indices i_α in the sum $\sum''$. Thus

$$A_n(r_1, \ldots, r_u) = \frac{(k_n)_u}{s_n^r} E[X_{n1}^{r_1} \cdots X_{nu}^{r_u}].$$

The problem again is to prove (30.9).

The proof goes by induction on u. Now $A_n(r) = k_n s_n^{-r} n^{-1} \sum_{h=1}^{n} x_{nh}^r$, so that $A_n(1) = 0$ and $A_n(2) = 1$. If $r \geq 3$, then $|x_{nh}^r| \leq M_n^{r-2} x_{nh}^2$, and so $|A_n(r)| \leq (M_n/s_n)^{r-2} \to 0$ by (30.11).

Next suppose as induction hypothesis that (30.9) holds with $u-1$ in place of u. Since the sampling is without replacement, $E[X_{n1}^r \cdots X_{nu}^{r_u}] = \sum x_{nh_1}^{r_1} \cdots x_{nh_u}^{r_u}/(n)_u$, where the summation extends over the u-tuples $(h_1, \ldots, h_u)$ of distinct integers in the range $1 \leq h_\alpha \leq n$. In this last sum enlarge the range by requiring of $(h_1, h_2, \ldots, h_u)$ only that $h_2, \ldots, h_u$ be distinct, and then compensate by subtracting away the terms where $h_1 = h_2$, where $h_1 = h_3$, and so on. The result is

$$E[X_{n1}^{r_1} \cdots X_{nu}^{r_u}] = \frac{n(n)_{u-1}}{(n)_u} E[X_{n1}^{r_1}] E[X_{n2}^{r_2} \cdots X_{nu}^{r_u}] - \sum_{\alpha=2}^{u} \frac{(n)_{u-1}}{(n)_u} E[X_{n2}^{r_2} \cdots X_{n\alpha}^{r_1+r_\alpha} \cdots X_{nu}^{r_u}].$$

This takes the place of the factorization made possible in (30.8) by the assumed independence there. It gives

$$A_n(r_1, \ldots, r_u) = \frac{n}{n-u+1} \frac{k_n - u + 1}{k_n} A_n(r_1) A_n(r_2, \ldots, r_u) - \frac{k_n - u + 1}{n-u+1} \sum_{\alpha=2}^{u} A_n(r_2, \ldots, r_1 + r_\alpha, \ldots, r_u).$$

By the induction hypothesis the last sum is bounded, and the factor in front goes to 0 by (30.11). As for the first term on the right, the factor in front goes to 1. If $r_1 \neq 2$, then $A_n(r_1) \to 0$ and $A_n(r_2, \ldots, r_u)$ is bounded, and so $A_n(r_1, \ldots, r_u) \to 0$. The same holds by symmetry if $r_\alpha \neq 2$ for some α other than 1. If $r_1 = \cdots = r_u = 2$, then $A_n(r_1) = 1$, and $A_n(r_2, \ldots, r_u) \to 1$ by the induction hypothesis.

Thus (30.9) holds in all cases, and $S_n/s_n \Rightarrow N$ follows by the method of moments.

Application to Number Theory

Let $g(m)$ be the number of distinct prime factors of the integer m; for example $g(3^4 \times 5^2) = 2$. Since there are infinitely many primes, $g(m)$ is unbounded above; for the same reason, it drops back to 1 for infinitely many m (for the primes and their powers). Since g fluctuates in an irregular way, it is natural to inquire into its average behavior.

On the space Ω of positive integers, let P_n be the probability measure that places mass $1/n$ at each of $1, 2, \ldots, n$, so that among the first n positive

integers the proportion that are contained in a given set A is just $P_n(A)$. The problem is to study $P_n[m: g(m) \le x]$ for large n.

If $\delta_p(m)$ is 1 or 0 according as the prime p divides m or not, then

$$g(m) = \sum_p \delta_p(m). \tag{30.12}$$

Probability theory can be used to investigate this sum because under P_n the $\delta_p(m)$ behave somewhat like independent random variables. If $p_1, \ldots, p_u$ are distinct primes, then by the fundamental theorem of arithmetic, $\delta_{p_1}(m) = \cdots = \delta_{p_u}(m) = 1$—that is, each p_i divides m—if and only if the product $p_1 \cdots p_u$ divides m. The probability under P_n of this is just n^{-1} times the number of m in the range $1 \le m \le n$ that are multiples of $p_1 \cdots p_u$, and this number is the integer part of $n/p_1 \cdots p_u$. Thus

$$P_n\bigl[m: \delta_{p_i}(m) = 1, i = 1, \ldots, u\bigr] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \cdots p_u} \right\rfloor \tag{30.13}$$

for distinct p_i.

Now let X_p be independent random variables (on some probability space, one variable for each prime p) satisfying

$$P\bigl[X_p = 1\bigr] = \frac{1}{p}, \qquad P\bigl[X_p = 0\bigr] = 1 - \frac{1}{p}.$$

If $p_1, \ldots, p_u$ are distinct, then

$$P\bigl[X_{p_i} = 1, i = 1, \ldots, u\bigr] = \frac{1}{p_1 \cdots p_u}. \tag{30.14}$$

For fixed $p_1, \ldots, p_u$, (30.13) converges to (30.14) as $n \to \infty$. Thus the behavior of the X_p can serve as a guide to that of the $\delta_p(m)$. If $m \le n$, (30.12) is $\sum_{p \le n} \delta_p(m)$, because no prime exceeding m can divide it. The idea[†] is to compare this sum with the corresponding sum $\sum_{p \le n} X_p$.

This will require from number theory the elementary estimate[‡]

$$\sum_{p \le x} \frac{1}{p} = \log\log x + O(1). \tag{30.15}$$

The mean and variance of $\sum_{p \le n} X_p$ are $\sum_{p \le n} p^{-1}$ and $\sum_{p \le n} p^{-1}(1 - p^{-1})$; since $\sum_p p^{-2}$ converges, each of these two sums is asymptotically $\log\log n$.

[†]Compare Problems 2.18, 5.19, and 6.16.

[‡]See, for example, Problem 18.17, or HARDY & WRIGHT, Chapter XXII.

Comparing $\sum_{p \le n} \delta_p(m)$ with $\sum_{p \le n} X_p$ then leads one to conjecture the *Erdös–Kac central limit theorem for the prime divisor function*:

Theorem 30.3. *For all x,*

$$P_n\left[m: \frac{g(m) - \log\log n}{\sqrt{\log\log n}} \le x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du. \tag{30.16}$$

PROOF. The argument uses the method of moments. The first step is to show that (30.16) is unaffected if the range of p in (30.12) is further restricted. Let $\{\alpha_n\}$ be a sequence going to infinity slowly enough that

$$\frac{\log \alpha_n}{\log n} \to 0 \tag{30.17}$$

but fast enough that

$$\sum_{\alpha_n < p \le n} \frac{1}{p} = o(\log\log n)^{1/2}. \tag{30.18}$$

Because of (30.15), these two requirements are met if, for example, $\log \alpha_n = (\log n)/\log\log n$.

Now define

$$g_n(m) = \sum_{p \le \alpha_n} \delta_p(m). \tag{30.19}$$

For a function f of positive integers, let

$$E_n[f] = n^{-1} \sum_{m=1}^{n} f(m)$$

denote its expected value computed with respect to P_n. By (30.13) for $u = 1$,

$$E_n\left[\sum_{p > \alpha_n} \delta_p\right] = \sum_{\alpha_n < p \le n} P_n[m: \delta_p(m) = 1] \le \sum_{\alpha_n < p \le n} \frac{1}{p}.$$

By (30.18) and Markov's inequality,

$$P_n\left[m: |g(m) - g_n(m)| \ge \epsilon(\log\log n)^{1/2}\right] \to 0.$$

Therefore (Theorem 25.4), (30.16) is unaffected if $g_n(m)$ is substituted for $g(m)$.

Now compare (30.19) with the corresponding sum $S_n = \sum_{p \le \alpha_n} X_p$. The mean and variance of S_n are

$$c_n = \sum_{p \le \alpha_n} \frac{1}{p}, \qquad s_n^2 = \sum_{p \le \alpha_n} \frac{1}{p}\left(1 - \frac{1}{p}\right),$$

and each is $\log\log n + o(\log\log n)^{1/2}$ by (30.18). Thus (see Example 25.8), (30.16) with $g(m)$ replaced as above is equivalent to

$$P_n\left[m: \frac{g_n(m) - c_n}{s_n} \le x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du. \tag{30.20}$$

It therefore suffices to prove (30.20).

Since the X_p are bounded, the analysis of the moments (30.7) applies here. The only difference is that the summands in S_n are indexed not by the integers k in the range $k \le k_n$ but by the primes p in the range $p \le \alpha_n$; also, X_p must be replaced by $X_p - p^{-1}$ to center it. Thus the rth moment of $(S_n - c_n)/s_n$ converges to that of the normal distribution, and so (30.20) and (30.16) will follow by the method of moments if it is shown that as $n \to \infty$,

$$E\left[\left(\frac{S_n - c_n}{s_n}\right)^r\right] - E_n\left[\left(\frac{g_n - c_n}{s_n}\right)^r\right] \to 0 \tag{30.21}$$

for each r.

Now $E[S_n^r]$ is the sum

$$\sum_{u=1}^{r} \sum{}' \frac{r!}{r_1! \cdots r_u!} \frac{1}{u!} \sum{}'' E\left[X_{p_1}^{r_1} \cdots X_{p_u}^{r_u}\right], \tag{30.22}$$

where the range of Σ' is as in (30.6) and (30.7), and Σ'' extends over the u-tuples $(p_1, \ldots, p_u)$ of distinct primes not exceeding α_n. Since X_p assumes only the values 0 and 1, from the independence of the X_p and the fact that the p_i are distinct, it follows that the summand in (30.22) is

$$E\left[X_{p_1} \cdots X_{p_u}\right] = \frac{1}{p_1 \cdots p_u}. \tag{30.23}$$

By the definition (30.19), $E_n[g_n^r]$ is just (30.22) with the summand replaced by $E_n[\delta_{p_1}^{r_1} \cdots \delta_{p_u}^{r_u}]$. Since $\delta_p(m)$ assumes only the values 0 and 1, from (30.13) and the fact that the p_i are distinct, it follows that this summand is

$$E_n\left[\delta_{p_1} \cdots \delta_{p_u}\right] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \cdots p_u} \right\rfloor. \tag{30.24}$$

But (30.23) and (30.24) differ by at most $1/n$, and hence $E[S_n^r]$ and $E_n[g_n^r]$ differ by at most the sum (30.22) with the summand replaced by $1/n$. Therefore,

$$|E[S_n^r] - E_n[g_n^r]| \le \frac{1}{n}\Bigl(\sum_{p \le \alpha_n} 1\Bigr)^r \le \frac{\alpha_n^r}{n}. \tag{30.25}$$

Now

$$E\bigl[(S_n - c_n)^r\bigr] = \sum_{k=0}^{r} \binom{r}{k} E[S_n^k](-c_n)^{r-k},$$

and $E_n[(g_n - c_n)^r]$ has the analogous expansion. Comparing the two expansions term for term and applying (30.25) shows that

$$\begin{aligned} &\bigl|E\bigl[(S_n - c_n)^r\bigr] - E_n\bigl[(g_n - c_n)^r\bigr]\bigr| \\ &\qquad \le \sum_{k=0}^{r} \binom{r}{k} \frac{\alpha_n^k}{n} c_n^{r-k} = \frac{1}{n}(\alpha_n + c_n)^r. \end{aligned} \tag{30.26}$$

Since $c_n \le \alpha_n$, and since $\alpha_n^r/n \to 0$ by (30.17), (30.21) follows as required. ■

The method of proof requires passing from (30.12) to (30.19). Without this, the α_n on the right in (30.26) would instead be n, and it would not follow that the difference on the left goes to 0; hence the truncation (30.19) for an α_n small enough to satisfy (30.17). On the other hand, α_n must be large enough to satisfy (30.18), in order that the truncation leave (30.16) unaffected.

PROBLEMS

30.1. From the central limit theorem under the assumption (30.5) get the full Lindeberg theorem by a truncation argument.

30.2. For a sample of size k_n with replacement from a population of size n, the probability of no duplicates is $\prod_{j=0}^{k_n-1}(1 - j/n)$. Under the assumption $k_n/\sqrt{n} \to 0$ in addition to (30.10), deduce the asymptotic normality of S_n by a reduction to the independent case.

30.3. By adapting the proof of (21.24), show that the moment generating function of μ in an arbitrary interval determines μ.

30.4. 25.13 30.3↑ Suppose that the moment generating function of μ_n converges to that of μ in some interval. Show that $\mu_n \Rightarrow \mu$.

30.5. Let μ be a probability measure on R^k for which $\int_{R^k}|x_i|^r\mu(dx)<\infty$ for $i=1,\dots,k$ and $r=1,2,\dots$. Consider the cross moments

$$\alpha(r_1,\dots,r_k)=\int_{R^k}x_1^{r_1}\cdots x_k^{r_k}\mu(dx)$$

for nonnegative integers r_i.

(a) Suppose for each i that

$$\sum_r \frac{\theta^r}{r!}\int_{R^k}|x_i|^r\mu(dx) \tag{30.27}$$

has a positive radius of convergence as a power series in θ. Show that μ is determined by its moments in the sense that, if a probability measure ν satisfies $\alpha(r_1,\dots,r_k)=\int x_1^{r_1}\cdots x_k^{r_k}\nu(dx)$ for all $r_1,\dots,r_k$, then ν coincides with μ.

(b) Show that a k-dimensional normal distribution is determined by its moments.

30.6. ↑ Let μ_n and μ be probability measures on R^k. Suppose that for each i, (30.27) has a positive radius of convergence. Suppose that

$$\int_{R^k}x_1^{r_1}\cdots x_k^{r_k}\mu_n(dx)\to\int_{R^k}x_1^{r_1}\cdots x_k^{r_k}\mu(dx)$$

for all nonnegative integers $r_1,\dots,r_k$. Show that $\mu_n\Rightarrow\mu$.

30.7. 30.5↑ Suppose that X and Y are bounded random variables and that X^m and Y^n are uncorrelated for $m,n=1,2,\dots$. Show that X and Y are independent.

30.8. 26.17 30.6↑ (a) In the notation (26.32), show for $\lambda\neq 0$ that

$$M\left[(\cos\lambda x)^r\right]=\binom{r}{r/2}\frac{1}{2^r} \tag{30.28}$$

for even r and that the mean is 0 for odd r. It follows by the method of moments that $\cos\lambda x$ has a distribution in the sense of (25.18), and in fact of course the relative measure is

$$\rho[x:\cos\lambda x\le u]=1-\frac{1}{\pi}\arccos u,\qquad -1<u<1. \tag{30.29}$$

(b) Suppose that $\lambda_1,\lambda_2,\dots$ are linearly independent over the field of rationals in the sense that, if $n_1\lambda_1+\cdots+n_m\lambda_m=0$ for integers n_ν, then $n_1=\cdots=n_m=0$. Show that

$$M\left[\prod_{\nu=1}^k(\cos\lambda_\nu x)^{r_\nu}\right]=\prod_{\nu=1}^k M\left[(\cos\lambda_\nu x)^{r_\nu}\right] \tag{30.30}$$

for nonnegative integers $r_1,\dots,r_k$.

(c) Let $X_1, X_2, \ldots$ be independent and have the distribution function on the right in (30.29). Show that

$$\rho\left[x: \sum_{j=1}^{k} \cos \lambda_j x \le u\right] = P[X_1 + \cdots + X_k \le u]. \tag{30.31}$$

(d) Show that

$$\lim_{k\to\infty} \rho\left[x: u_1 < \sqrt{\frac{2}{k}} \sum_{j=1}^{k} \cos \lambda_j x \le u_2\right] = \frac{1}{\sqrt{2\pi}} \int_{u_1}^{u_2} e^{-v^2/2}\, dv. \tag{30.32}$$

For a signal that is the sum of a large number of pure cosine signals with incommensurable frequencies, (30.32) describes the relative amount of time the signal is between u_1 and u_2.

30.9. 6.16↑ From (30.16), deduce once more the Hardy–Ramanujan theorem (see (6.10)).

30.10. ↑ (a) Prove that (if P_n puts probability $1/n$ at $1, \ldots, n$)

$$\lim_n P_n\left[m: \left|\frac{\log\log m - \log\log n}{\sqrt{\log\log n}}\right| \ge \epsilon\right] = 0. \tag{30.33}$$

(b) From (30.16) deduce that (see (2.35) for the notation)

$$D\left[m: \frac{g(m) - \log\log m}{\sqrt{\log\log m}} \le x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du. \tag{30.34}$$

30.11. ↑ Let $G(m)$ be the number of prime factors in m with multiplicity counted. In the notation of Problem 5.19, $G(m) = \sum_p \alpha_p(m)$.

(a) Show for $k \ge 1$ that $P_n[m: \alpha_p(m) - \delta_p(m) \ge k] \le 1/p^{k+1}$; hence $E_n[\alpha_p - \delta_p] \le 2/p^2$.

(b) Show that $E_n[G - g]$ is bounded.

(c) Deduce from (30.16) that

$$P_n\left[m: \frac{G(m) - \log\log n}{\sqrt{\log\log n}} \le x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

(d) Prove for G the analogue of (30.34).

30.12. ↑ Prove the Hardy–Ramanujan theorem in the form

$$D\left[m: \left|\frac{g(m)}{\log\log m} - 1\right| \ge \epsilon\right] = 0.$$

Prove this with G in place of g.